Identifying Calcium-Binding Sites and Predicting Disulfide Connectivity

Hai Deng

Follow this and additional works at: https://scholarworks.gsu.edu/cs_diss

Recommended Citation

This Dissertation is brought to you for free and open access by the Department of Computer Science at ScholarWorks @ Georgia State University. It has been accepted for inclusion in Computer Science Dissertations by an authorized administrator of ScholarWorks @ Georgia State University. For more information, please contact scholarworks@gsu.edu.
IDENTIFYING CALCIUM-BINDING SITES AND PREDICTING DISULFIDE CONNECTIVITY

by

HAI DENG

Under the Direction of Guantao Chen

ABSTRACT

Most questions in proteomics require complex answers. Yet graph theory, supervised learning, and statistical model have decomposed complex questions into simple questions with simple answers. The expertise in the field of protein study often address tasks that demand answers as complex as the questions. Such complex answers may consist of multiple factors that must be weighed against each other to arrive at a globally satisfactory and consistent solution to the question.

In the prediction of calcium binding in proteins, we construct a global oxygen contact graph of a protein, then apply a graph algorithm to find oxygen clusters with the fixed size of four, finally employ a geometry algorithm to judge if the oxygen clusters are calcium-binding sites or not. Additionally, we can predict the locations of those sites.

Furthermore, we construct a global oxygen contact graph including oxygen-bonded carbon atoms of a protein, then apply a graph algorithm to find local biggest oxygen clusters, finally design another geometric filter to exclude the non-calcium binding oxygen clusters. In addition, we apply observed chemical properties as a chemical filter to recognize some non-calcium binding oxygen clusters.

In order to explore the characteristics of calcium-binding sites in proteins, we conduct a statistic survey on four datasets derived from 1994 to 2005 about the geometric parameters and chemical properties of calcium-binding sites.

In the prediction of disulfide bond connectivity, we analyze protein sequences to predict the folding of proteins relative to the cystines using nearest neighboring methods. we extend a new pattern-wise method to all available template proteins, and find global pattern of pairing cysteines with a new descriptor of cysteine separation profile on protein secondary structure.

INDEX WORDS: Calcium-binding sites, Disulfide connectivity, Graph algorithm, Geometrical structure, Nearest neighboring methods
IDENTIFYING CALCIUM-BINDING SITES AND PREDICTING DISULFIDE CONNECTIVITY

by

HAI DENG

A Dissertation Submitted in Partial Fulfillment of the Requirements for the Degree of

Doctor of Philosophy

in the College of Arts and Sciences

Georgia State University

2007
IDENTIFYING CALCIUM-BINDING SITES AND PREDICTING DISULFIDE CONNECTIVITY

by

HAI DENG

Major Professor: Guantao Chen
Committee: Yi Pan
Rajshekhar Sunderraman
Jenny J. Yang

Electronic Version Approved:

Office of Graduate Studies
College of Arts and Sciences
Georgia State University
August 2007
ACKNOWLEDGMENTS

First and foremost, I would like to thank my advisor, Professor Guantao Chen, for all his help, advice, support, guidance, and patience. He always helped me whenever I needed intelligent knowledge. He always encouraged me whenever I was frustrated. He always inspired me to strive for the best. I have been very lucky to work with him.

I am grateful to Dr. Yi Pan, Dr. Rajshekhar Sunderraman, and Dr. Jenny J. Yang for serving on my Ph.D committee, and for their time and co-operation in reviewing this work. Dr. Pan helps me a lot for academic advice on publication in computer science community. I would like to thank Dr. Sunderraman specially for providing unselfish help during my difficult situation in the academic life. Special thanks also go to Dr. Yang and all members of her lab for their expertise on biochemistry field.

Last but not least, I would like to thank my family for their support and encouragement for me to pursue a Ph.D. in the United States. Specially, my parents are over seventy and have never gave up the hope of my finishing Ph.D. degree. Although they have not seen me during my Ph.D. study, they are always asking me about my study by email and phone. It is their consistent care to make me finish this dissertation. I appreciate and love them in all my life.
# TABLE OF CONTENTS

Acknowledgments ........................................ iv  
List of Tables ........................................ vii  
List of Figures ........................................ viii  
List of Acronyms ...................................... xi  

Chapter 1: Introduction ................................ 1  
1.1 Motivation ........................................ 1  
1.2 Problem description and related works ........ 2  
1.3 Contributions .................................... 8  
1.4 Organization .................................... 9  

Chapter 2: The graph and geometry approach (GG) for predicting calcium binding .......... 11  
2.1 Introduction ...................................... 11  
2.2 Methods ......................................... 13  
2.3 Results and Discussions ......................... 16  

Chapter 3: Identifying Calcium-binding Sites with Geometric and Chemical Criteria at Oxygen-Carbon Shell .......... 28  
3.1 Introduction ...................................... 28  
3.2 Methods ......................................... 30  
3.3 Results and Discussions ......................... 36
<table>
<thead>
<tr>
<th>Chapter 4: Feature Analysis on Calcium-binding Sites</th>
<th>41</th>
</tr>
</thead>
<tbody>
<tr>
<td>4.1 Introduction</td>
<td>41</td>
</tr>
<tr>
<td>4.2 Method</td>
<td>43</td>
</tr>
<tr>
<td>4.3 Results</td>
<td>45</td>
</tr>
<tr>
<td>Chapter 5: Predicting Disulfide Connectivity</td>
<td>59</td>
</tr>
<tr>
<td>5.1 Introduction</td>
<td>59</td>
</tr>
<tr>
<td>5.2 Method</td>
<td>60</td>
</tr>
<tr>
<td>5.3 Results</td>
<td>64</td>
</tr>
<tr>
<td>5.4 Conclusion and Discussion</td>
<td>67</td>
</tr>
<tr>
<td>Chapter 6: Conclusions</td>
<td>71</td>
</tr>
<tr>
<td>Bibliography</td>
<td>72</td>
</tr>
<tr>
<td>Table Number</td>
<td>Description</td>
</tr>
<tr>
<td>--------------</td>
<td>-----------------------------------------------------------------------------</td>
</tr>
<tr>
<td>2.1</td>
<td>Calcium coordination numbers (CN, $Ca - O \leq 3.5\text{A}$) in three datasets</td>
</tr>
<tr>
<td>4.1</td>
<td>Coordination numbers (CN) of four datasets</td>
</tr>
<tr>
<td>4.2</td>
<td>Ligand Charge Data for Two Calcium Binding Sites</td>
</tr>
<tr>
<td>5.1</td>
<td>Performance comparison among different algorithms</td>
</tr>
</tbody>
</table>
## LIST OF FIGURES

<table>
<thead>
<tr>
<th>Figure Number</th>
<th>Description</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>1.1</td>
<td>The binding affinities of calcium to calcium-binding proteins are related to the calcium concentrations in the biological environment, such as extracellular, intracellular compartments, and the nucleus [14, 83]. The dissociation constants (Kd) of the regulated proteins are in the ranges of the calcium concentrations where the proteins are located.</td>
<td>3</td>
</tr>
<tr>
<td>1.2</td>
<td>The formation of a disulfide bond</td>
<td>7</td>
</tr>
<tr>
<td>2.1</td>
<td>The schematic model of GG program is shown. The positions of oxygen atoms (dark dots) are extracted from the protein structure while the other atoms (light dots) are excluded. The distance between two oxygen atoms are calculated and an edge is assigned if the distance is below a cutoff distance (O-O cutoff). A potential calcium-binding position is a clique (right bottom), which is a group in which every oxygen atom is linked to all other members of the group by edges.</td>
<td>14</td>
</tr>
<tr>
<td>2.2</td>
<td>The pseudo code of the algorithm for finding all cliques with size four</td>
<td>15</td>
</tr>
<tr>
<td>2.3</td>
<td>The calmodulin (A, 3CLN), galactose-binding protein (B, 1GCG), flavodoxin (C, 1AG9), and serum amyloid P component (D, 1SAC) represent different classes of calcium-binding sites including different ligand distributions (continuous, semi-continuous, and discontinuous), binding numbers (single site, independent multiple sites, and sites with shared ligands), cofactor conditions, and protein sizes. The green balls are calcium ions. The ligand oxygen atoms are from proteins (red balls) and cofactors (orange balls). The protein frames are in light blue while the cofactors are in orange. The pictures are generated by PyMol (DeLano Scientific).</td>
<td>18</td>
</tr>
</tbody>
</table>
2.4 The site sensitivity (the true predicted sites in the total sites) as a function of O-O distance cutoff ......................................................... 20

2.5 The sensitivity (black bars) and selectivity (grey bars) of the GG program with the O-O distance cutoff of 6.0 Å and the psdCa-O distance range cutoff of 1.8-3.0 Å. The standard of a true prediction is that the predicted calcium position is within 3.5 Å(A) or 1.0 Å(B) threshold distance to the documented calcium. The results of all sites are shown on the left and the results only considering the sites with 4 or more ligands are shown on the right. .......................................................... 22

3.1 The schematic model of GG2.0 Program is shown. The positions of oxygen atoms in green (non-water oxygen) and orange (water oxygen) and bonded carbon atoms in blue are extracted from the protein structure. An edge is given between two oxygen atoms if the distance between them is less than the cutoff distance. A putative calcium-binding site is an oxygen cluster/clique with a constraint on the average radius of two clusters (right bottom). .................................................. 31

3.2 The procedure of finding the maximal cliques. .................................................. 33

3.3 The histogram of RO and RC (a), the histogram of bidentate r_RO_RC, bidentate r_RO_RC and r_RO_RC (b), and the histogram of bidentate ar_RO_RC, bidentate ar_RO_RC and ar_RO_RC (c). .................................................. 37

3.4 The effect of ar_RO_RC on prediction accuracy .................................................. 38

3.5 The Correlation of r_RO_RC and avg Angle of Ca-O-C .................................. 40

4.1 Illustration of two angles: Ca-O-C and O-Ca-O, and two distances: dist(Ca,C) and dist(C,O) .......................................................... 46

4.2 Distribution of average Ca-O (Avg Ca-O) and average Ca-C (Avg Ca-C) distances . 48

4.3 The distributions of r_RO_RC (a) and ar_RO_RC (b) in dataset I (light gray), dataset II (dark gray), dataset III (white), and dataset IV (black) ................. 49

4.4 Distribution of r_RO_RC for non-bidentate (black) and bidentate sites (white) .... 50
4.5 Distribution of Ca-O-C angles for bidentate and monodentate residues over four datasets .................................................. 51
4.6 Distribution of Ca-O-C angle pair sums for bidentate residues over four datasets ............................................. 52
4.7 Distribution of Ca-O-C angle pair sums for bidentate residues over four datasets ............................................. 53
4.8 Frequency of occurrences of formally charged residues by site, for Datasets I-IV. A formal charge of -3 was most frequent for the sum of all four datasets, however, a formal charge of 2 was disproportionately high within Dataset II .................................................. 54
4.9 A) 1BJR calcium 290 has a formal charge of zero in the binding site, and B) 1HYT, calcium 801, has a formal charge of 5 .................................................................................................. 56
4.10 Pentagonal-bipyrimid geometry surrounding Ca$^{2+}$ ion (A) and (B) ligand-ion contact geometry for calmodulin EF-I binding pocket from PDB 3CLN. The water molecule at -X is not shown. .................................................. 58

5.1 An example of finding the pattern of the test protein (a) from the subpattern of the template protein (b) with more disulfide bonds .............................................................................. 68
5.2 An example of finding the pattern of the test protein (a) from the template protein (b) with less disulfide bonds .............................................................................. 69
LIST OF ACRONYMS

1NN: One nearest neighboring method

CSPSS: Cysteine separations profiles on secondary structure

GG: Graph and geometry approach
Chapter 1

INTRODUCTION

In the recent years, many biological and chemical problems search assistance from intensive computation with explosion of experimental data. It is most challenging in bioinformatics research that how to transform a biological or chemical problem into a math problem or a computational model.

In this dissertation, we choose two biochemical problems of identifying calcium binding and predicting disulfide connectivity in proteins as targets, employ graph representation to transform these two problems into computational problems, solve these problems with a series of methods such as graph-theoretic algorithm and one nearest neighboring method as well as statistical analysis of chemical properties.

1.1 Motivation

Proteomics is the large-scale study of proteins, particularly their structures and functions. This term was coined to make an analogy with genomics, and while it is often viewed as the "next step", proteomics is much more complicated than genomics. Most importantly, while the genome is a rather constant entity, the proteome differs from cell to cell and is constantly changing through its biochemical interactions with the genome and the environment. One organism will have radically different protein expression in different parts of its body, in different stages of its life cycle and in different environmental conditions. This technology is instrumental in biomarker discovery.

The entirety of proteins in existence in an organism throughout its life cycle, or on a smaller scale the entirety of proteins found in a particular cell type under a particular type of stimulation, are referred to as the proteome of the organism or cell type respectively. with completion of a rough draft of the human genome, many researchers are now looking at how genes and proteins interact
to form other proteins. A surprising finding of the Human Genome Project is that there are far fewer protein-coding genes in the human genome than there are proteins in the human proteome (22,000 genes vs. 400,000 proteins). The large increase in protein diversity is thought to be due to alternative splicing and post-translational modification of proteins. This discrepancy implies that protein diversity cannot be fully characterized by gene expression analysis alone, making proteomics a useful tool for characterizing cells and tissues of interest. To catalog all human proteins and ascertain their functions and interactions presents a daunting challenge for scientists. An international collaboration to achieve these goals is being coordinated by the Human Proteome Organisation.

However, experiments to determine protein structure and function are too expensive and time-consuming. Also it is impractical to analyze amounts of data manually. To help biologists and chemists reduce the workload, quite a few researchers in computer science and math apply data mining techniques to retrieve significant information from raw experimental data.

1.2 Problem description and related works

There are two problems of calcium binding and disulfide connectivity described as follows as well as with related works.

1.2.1 Calcium binding in proteins

Biological functions of calcium

Calcium is the fifth most abundant element in the earth’s crust and has no uncombined form in nature. Calcium, as a second messenger, regulates many cellular processes. However, all these regulating activities are not independent and relevant to the interaction with proteins. In different biological environments, calcium may have different affinities with binding proteins. [62, 87]. Calcium and calcium-binding proteins (CaBPs) play important impacts in intracellular signal transduction pathways and are associated with a wide-range of effects in health and disease [62, 85]. The binding affinities of calcium to calcium-binding proteins are related to the calcium concentrations in the biological environment, such as extracellular, intracellular compartments, and the nucleus [14, 83] (Fig. 1.1).
Figure 1.1: The binding affinities of calcium to calcium-binding proteins are related to the calcium concentrations in the biological environment, such as extracellular, intracellular compartments, and the nucleus [14, 83]. The dissociation constants (Kd) of the regulated proteins are in the ranges of the calcium concentrations where the proteins are located.

Abnormal calcium concentrations or abnormal calcium binding affinity in cells will lead to the loss of binding ability or improper binding, further to cause many diseases. For example, the loss of calcium-binding affinity of cadherin, an extracellular calcium-dependent cell adhesion molecule, will lead to the loss in adhesive function of the protein, which is often found in diffuse-type gastric cancers [8]. In addition, the over-expression of S100A6, an intracellular EF-hand protein reportedly deduce breast cancer through a calcium-dependent mechanism. In order to understand the role of calcium at the molecular level in signal transduction and disease, and thereby to establish further strategies for the development of therapeutic drugs, it is critic to understand what structural factors and amino acid properties control the calcium-binding affinity.

Related Important Works

1. Yamashita’s Method [107] Metal binding sites in proteins are varied in their coordination numbers and geometries, their metal preferences, and their ligands (which include backbone carbonyl oxygens; sidechain groups of aspartic, asparagine, cysteine, glutamic, glutamine, histidine, methionine, serine, threonine, and tyrosine residues; and water molecules). However, Yamashita et al. have found that metal sites of high hydrophobicity contrast from visual examination of structures of ionophores and metal-host molecules deposited in the Cambridge Structural Database. This visual impression of metals binding to centers of high
hydrophobicity contrast can be displayed graphically by a radial distribution plot.

The protein structure was embedded in a fine three-dimensional grid of points with a spacing 0.5 Å. Each potential calcium binding site was defined as a grid point located not within the van der Waals radius of a protein atom. The greedy algorithm searched all grid points satisfying the above condition. For every allowable grid point, a hydrophobicity contrast function was used to calculate its value. The hydrophobicity contrast function, a simple function of the distribution of the atoms within a molecule, gives a quantitative expression to the radial distribution. The contrast function has the desirable property that is generally a maximum when centered at or near a metal binding site. The hydrophobicity contrast function $C$ is evaluated for spheres of $R = 3.5 Å$ and $R = 7.0 Å$. $S$ scales the contrast value at 7.0 Å to the value at 3.5 Å. The simplest algorithm is to calculate looping over all grid points, finding all atoms within the probe radius $R$ of the grid point, and evaluating their contributions to the contrast function. However, this is extravagantly slow. A revised algorithm loops over the protein atoms, ignoring the metal ions. The contribution of $C$ of atoms to the grid points is added to those grid points. Then the program will loop through all points to found the greatest 1000 points and lowest 1000 points. The values have a clear cutoff between metal binding sites and non metal binding sites. The method was successful in Fe, Ca, and Cu but only in some cases and with low precision. They have noted the inefficiency of the algorithm and left that for later researchers.

2. Nayal and Di Cera’s Method [75]

Much in the spirit of Yamashita’s method, Nayal and Di Cera proposed to use valence function to calculate the potential calcium binding site’s local value. The protein structure is embedded in a very fine three-dimensional grid of points with a spacing of 0.1 Å. For each potential calcium-binding site, defined as a grid point not within the van der Waals radius of a protein atom increased by the ionic radius of calcium ion and surrounded by at least three oxygen atoms within a probe radius of 3.4Å, the value of the valence at some grid point is computed by summing the contribution of all oxygens according to valence function. In this equation, there are two empirical values varying with different metal ions. In constructing the algorithm,
they have found that looping over grid points would be very inefficient, an aspect already mentioned in Yamashita’s method. In practice, the algorithm loops over all atoms and ignores water molecules (water oxygen atoms were found to add considerable noise and very little information). For each oxygen atom, the allowable grid points within the probe radius are determined, and the contribution of the oxygen atom to the valence of each point is computed according to valence equation. All potential calcium-binding sites as defined above are stored as “pseudo-atoms” in a PDB-formatted output file. The file can be used to retrieve grid points to be superimposed to the protein structure for molecular graphics. The valences are then sorted for quantitative analysis.

They found that their method can obtain 94% site sensitivity and 95% site selectivity for a dataset of 32 proteins with 62 calcium binding sites. In addition, they improved the grid search algorithm in which another condition besides van der Waals radius was added, the potential calcium binding site must be surrounded by at least three oxygen atoms within 3.4 Å from it. Furthermore, their method loops over all atoms instead of grid points to find the qualified points. Finally, points to be calculated were sharply reduced. However, the computational time cost is still expensive. The calculation of site positive prediction leads to multiple redundancy, which means that there are bunches of predictions for one calcium binding site.

3. Fold-X Force Method [84]

The current state-of-the-art method to locate calcium binding sites is the Fold-X force method. The algorithm for the prediction of metal ions and water-binding sites in proteins is based on the method of Pitt and Goodfellow. First, the position of the ligand of interest calcium ion relative to the coordinating protein atoms is extracted from the high-resolution crystal structures. The result is a cloud of calcium ions around the protein-coordinating atom that can be divided into one or more centers using a clustering procedure. The clusters are then used as the canonical positions of interaction between the ligand and the protein atom. The metal-binding or water-binding-site algorithm in Fold-X starts by placing all canonical calcium ions on the potential coordinating atoms of a given protein structure. Calcium ions that clash
with the protein are removed, and the remaining ligands are searched for partial overlaps. Those ions that are closer than a 2.1Å are marked for fusion, i.e., the formation of one new ligand-binding position that replaces the canonical positions that gave rise to the fusion. The fusion only actually occurs when a search procedure starting from a position between the canonicals finds a new position that is within a predefined distance of both coordinating protein atom and has no van der Waals clashes with the other atoms in the structure. This procedure can be repeated, allowing fusion of more than two canonical ligands to form more complex binding sites, depending on the orbital properties of the ligand of interest. For calcium ions, they regulate at least four and at most six ligands for fusion. At the end of the fusion round, an optimization algorithm is used to find the optimal position of the predicted ligand according to the Fold-X free energy, and the resulting position is used to determine the binding energy of the ligand to the protein. Finally, we eliminate those predicted single atom ligands with low binding energy and hydrate them.

The greatness of the method is not only in the term of preciseness but also in its by-product that the method can solve the binding affinity problem. Instead of using grid search algorithm, the method fuses the possible metal ions from the solvation model of Fold-X. At the end of fusion, an optimization algorithm is used to find the optimal position for the binding site using the Fold-X free energy function. The optimization algorithm here is still local grid searching thus inefficient in time. In addition, this method needs high resolution coordinated atoms.

1.2.2 Disulfide Connectivity in Proteins

Disulfide Bond

A disulfide bond (SS-bond) [1, 67], also called a disulfide bridge, is a strong covalent bond between two sulphydryl groups. This bond is very important to the folding, structure, and function of proteins.

When two amino acids covalently bond to each other through their side chains, they normally do so through a disulfide bond. The particular side chain involved is the sulphydryl group (−SH). Oxidation of the thiol group yields a disulfide (S-S) bond. The presence of S-S then helps to maintain the tertiary structure of the protein. An amino acid that commonly forms S-S bonds in
proteins is cysteine. When two cysteines are bonded by an S-S bond, the resulting molecule between the two protein chains is called cystine. The figure below shows the formation of a disulfide bond. The R on each side group represents the remainder of the amino acid. In proteins that contain more than one disulfide bond, proper pairing of the cysteine residues is important for normal structure and activity. Disulfide bonding patterns are more important.

**Related Important Works**

1. CSP [111]

Zhao et al. adopt cysteine separation profiles (CSPs) to predict the disulfide connectivity of proteins. The separations among oxidized cysteine residues on a protein sequence have been encoded into vectors named cysteine separation profiles (CSPs). Through comparisons of their CSPs, the disulfide connectivity of a test protein is inferred from a non-redundant template set. For non-redundant proteins in SwissProt 39 (SP39) sharing less than 30% sequence identity, the prediction accuracy of a four-fold cross validation is 49%. The prediction accuracy of disulfide connectivity for proteins in SwissProt 43 (SP43) is even higher (53%). The method proposed in this work is relatively simple and can generate higher accuracies comparing to conventional methods.

2. SVM [97]

Tsai et al. is the first team to use support vector machine (SVM) for predicting disulfide
connectivity. In this study, a descriptor derived from the sequential distance between oxidized cysteines (denoted as DOC) is proposed. This approach based on weighted graph matching was further developed to predict the disulfide connectivity pattern in proteins. When DOC was applied, prediction accuracy of 63% for our SVM models could be achieved, which is significantly higher than those obtained from previous approaches. The results show that using the non-local descriptor DOC coupled with local sequence profiles significantly improves the prediction accuracy. These improvements demonstrate that DOC, with a proper scaling scheme, is an effective feature for the prediction of disulfide connectivity.

3. Two-level Model [20]

Chen et al. combine the previous two methods into two level models of SVM detailed as following:

level-1 models: pair-wise SVM

the information generated in the pair-wise SVM models to further refine the prediction of disulfide connectivity. That is, using the pair-wise SVM, the bonding probability for each cysteine pairs which will be used as inputs in the level-2 models.

level-2 models: pattern-wise models

In the second level, for a protein, all the possible bonding patterns of disulfide connectivity are considered. Each bonding pattern is encoded with the following features:

(1) Confidence score from level-1 (S): according to the bonding probability, they compute a score for the bonding pattern. (2) CSP search result (C): using CSP, they examine if the pattern is the same as the one searched by CSP. (3) Global information of the protein (G): they further include protein information, such as protein length, for the input feature.

They have made a breakthrough to increase the pattern accuracy to 70%.

1.3 Contributions

Here, we give the overall contributions. They are basically summarized as follows:
1. Graph representation of protein

Each unit of a protein can be taken as a vertex in the graph. For instance, Vertices represent oxygen atoms in constructed protein graphs for calcium binding problem whereas vertices represent cysteine residues for disulfide bonding connectivity problem. Edges can indicate any interacting relation between vertices. They are either non-weighted like spatial constraints in the oxygen graph in the calcium binding problem or weighted like contact potential in the disulfide bonding problem. We use graph construction to derive a simplified and selected feature structure from a whole protein.

2. Integration of graph algorithms and geometrical algorithms

Two different clique-finding graph algorithms are employed with two different geometrical algorithms to identify calcium-binding sites. These approaches are hybrid systems integrating different algorithms.

3. Statistical analysis of calcium-binding sites

To explore the secrets and help later prediction of calcium-binding sites, we adopt four datasets as samples to conduct a survey of geometrical, chemical and biophysical characteristics of calcium-binding sites.

4. New descriptor and method for predicting disulfide connectivity

CSPSS is applied to the disulfide bond pattern problem to calculate the difference between any two proteins with the same number of known half cystines. Meanwhile we develop a new pattern-wise method to calculate the difference between any two proteins even with the different number of known half cystines. Both descriptor and method are available for hybrid models to predict disulfide connectivity.

1.4 Organization

Chapter 2 detailed the Graph and geometry approach (GG) to predict the calcium binding location given a protein structure. Chapter 3 explains how to design geometric and chemical filter to exclude
non-calcium binding clusters on the same problem as that in Chapter 2. Chapter 4 conducts statistical analysis of characteristics of calcium-binding sites. Chapter 5 introduces the new descriptor, CSPSS, as well as new pattern-wise method to predict disulfide connectivity in proteins. Chapter 6 concludes the existing work and propose future work.
Chapter 2

THE GRAPH AND GEOMETRY APPROACH (GG) FOR PREDICTING CALCIUM BINDING

Identifying calcium-binding sites in proteins is one of the first steps towards predicting and understanding the role of calcium in biological systems for protein structure and function studies. Due to the complexity and irregularity of calcium-binding sites, a fast and accurate method for predicting and identifying calcium-binding protein is needed. Here we report our development of a new fast algorithm (GG) to detect calcium-binding sites. The GG algorithm uses a graph theory algorithm to find oxygen clusters of the protein and a geometric algorithm to identify the center of these clusters. A cluster of four or more oxygen atoms has a high potential for calcium binding. High performance with about 90% site sensitivity and 80% site selectivity has been obtained for three datasets containing a total of 123 proteins. The results suggest that a sphere of a certain size with four or more oxygen atoms on the surface and without other atoms inside is necessary and sufficient for quickly identifying the majority of the calcium-binding sites with high accuracy. Our finding opens a new avenue to visualize and analyze calcium-binding sites in proteins facilitating the prediction of functions from structural genomic information.

2.1 Introduction

Calcium regulates many biological processes through its interactions with numerous calcium-binding proteins [51, 53]. In addition to stabilizing the proteins, calcium also induces conformational changes to switch on and off the biological functions [50, 76]. Calcium ions are predominantly chelated by protein oxygen atoms from carboxyl sidechain of Asp, Glu, Asn and Gln, hydroxyl group from Ser and Thr and carbonyl main of the protein. In addition, it can be chelated by oxygen atoms from solvent water, phosphate, carbohydrate and lipids [72, 75]. Although the coordination number of calcium varies from 3 to more than 10 in small molecules, it is typically 5-8 in proteins with an average of about 6.5 - 7 [77, 109]. Our studies and others showed that most of the
calcium-oxygen distances in proteins vary from 2 to 3 Å with an average about 2.4 Å and different classes of calcium-binding sites in proteins can be identified using a pentagonal bipyramidal geometry with Ca-O bond lengths of 2.4 ± 1.0 Å and common calcium binding ligand residues [77]. This finding has facilitated us to apply geometry-based algorithm in addition to the charge and chemical properties to design calcium-binding proteins with biological functions [109, 110, 108]. There is a strong need in developing methodology to predict and visualize calcium-binding sites in proteins with high speed. Due to the rapid progress in NMR techniques, the solution structures of proteins are significantly increased [9]. Unfortunately, NMR cannot directly provide the calcium-binding coordination. While X-ray crystallography has been the major tool to visualize calcium-binding sites, calcium-binding sites with weak affinities are less defined if not completed unknown. For example, although it is known that calcium is essential for the function of metabotropic glutamate receptors (mGluR), no calcium-binding sites were observed in several X-ray structures of the extracellular domains of the mGluR [34, 57, 98]. In addition, as the worldwide development in structural genomics initiative speeds up to solve numerous protein structures, the prediction of protein’s functions and metal-binding properties becomes more and more important [11, 13, 58, 59, 103, 113]. Identifying calcium-binding sites is not only crucial for the study of individual proteins but also helpful for revealing the general factors involved in such as the mechanisms governing calcium-binding affinity, selectivity, and calcium-induced conformational change. A fast and accurate methodology for predicting calcium-binding sites will facilitate the understanding and prediction of functions of calcium roles in biological systems (denoted as calciomics) [89, 104, 108]. Several methods, such as statistical approaches and neural network classifiers, have been used to identify the calcium-binding sites in proteins [5, 52, 7, 61, 60, 80, 89, 7]. Methods such as FEATURE, seqFEATURE, or MetSite involve a large number of properties of proteins and reveal a great deal of statistical significance between calcium- and non-calcium-binding sites. Eisenberg and coworkers observed that metal ions bind at centers of high hydrophobicity contrast in a number of cases [107]. Nayal and Di Cera improved performance through replacing the contrast function with valence function with a clear cutoff between calcium- and non-calcium-binding sites [75]. Both methods used grid algorithms to identifying potential calcium binding sites, which cause a huge amount of calculations. The computation complexities of these grid algorithms prevent their application to predicting calcium-binding sites in proteins with required speed and accuracy. To facilitate identifying, predicting, and
analyzing calcium-binding sites in proteins, we extend our geometric approach [109] by applying a
graph theory algorithm, taking advantage of its capability of extracting key features of the system
[15, 81].

2.2 Methods

2.2.1 Datasets

To compare our studies with previous findings and present different classes of calcium binding
proteins, three datasets were used. Nayal-Di Cera’s dataset (Dataset I) contains 32 proteins with
62 calcium-binding sites [75] and it was used for parameter adjustment. Liang’s dataset (Dataset
II) contains 54 proteins with 91 calcium-binding sites [61] and 14 non-calcium-binding proteins.
Dataset III from Pidcock and Moore contains 94 sites in 44 proteins representing all classes and
folds of calcium-binding proteins [77]. All structures are obtained from X-ray crystallography and
114 of these 123 structures have resolution < 2.4Å. Except for those from water, all of the oxygen
atoms including those from proteins, carbohydrates, lipids and other cofactors, are included in the
calculation.

2.2.2 Graph algorithm

For a given protein structure as shown in Fig. 2.1, the coordinates in the PDB file of oxygen
atoms including the heteroatoms but excluding those from water molecules were extracted first for
the analysis. The distances between every two oxygen atoms are calculated. A graph, \( G(V, E) \), is
constructed accordingly, in which each vertex in \( V \) represents an oxygen atom and each edge in \( E \)
represents a pair of oxygen atoms apart within an O-O cutoff distance. The graph construction
time is \( O(n^2) \), where \( n = |V(G)| \) is the number of oxygen atoms. A clique \( Q \), which represents
oxygen clusters, is a subset of \( V \) such that every two vertices of \( Q \) are adjacent.

The size of a clique refers to the number of vertices in the clique. Cliques may overlap represent-
ing multiple sites with shared ligand oxygens. We use a backtracking algorithm to find cliques of
certain sizes (4 in most of the studies reported here) in \( G \) (Fig. 2.2. The computational complexity,
\( Cost(G) \leq O(n\Delta^3) \), can be derived from the lines 1- 5 directly and \( \Delta \) is the maximum degree in
\( G \). For a general graph, \( \Delta \) can be as big as \( n-1 \) so that the computational complexity is \( O(n^4) \).
Figure 2.1: The schematic model of GG program is shown. The positions of oxygen atoms (dark dots) are extracted from the protein structure while the other atoms (light dots) are excluded. The distance between two oxygen atoms are calculated and an edge is assigned if the distance is below a cutoff distance (O-O cutoff). A potential calcium-binding position is a clique (right bottom), which is a group in which every oxygen atom is linked to all other members of the group by edges.
1. **Loop** through all vertices $a$ in $V$
2. **Loop** through all vertices $b$ in the adjacency list of $a$
3. output $V':=V'\{a\}$ if no such $b$ exists
4. **Loop** through all vertices $c$ adjacent to both $a$ and $b$
5. output $V:=V'\{a\}$ if no such $c$ exists
6. **Loop** through all vertices $d$ adjacent to both $a$, $b$, and $c$
7. output clique $\{a,b,c,d\}$
8. output $V:=V'\{a\}$ if no such $d$ exists

Figure 2.2: The pseudo code of the algorithm for finding all cliques with size four

However, in our case, $\Delta$ is bounded by a constant independent from $n$ due to the manner of graph construction. Therefore, $Cost(G) \leq O(n)$, which implies that the complexity of clique searching is linear to the total number of oxygen atoms.

2.2.3 **Geometry algorithm**

The circumcenter (CC) is defined to have the same distance to the four vertices of a clique (denoted as psdCa-O). A unique CC exists as long as four oxygens are not in one plane. To eliminate false positives, a clique is considered as a putative calcium-binding site only if psdCa-O falls into the range $(R1, R2)$, where $R1$ and $R2$ are the lower and upper limits. A D-filter is further applied to eliminate any cliques that contain non-oxygen atoms within a short distance (D-filter) from the CC since the space for calcium binding should not be occupied by other atoms. The complexity of this geometry algorithm is $O(nm)$, where $m$ is the number of all atoms other than oxygen.

2.2.4 **Removal of the redundant predictions**

To remove redundant predictions in one location, a merging algorithm was adopted. All putative binding sites in a protein are input in a vector and sorted by psdCa-O. The one with the shortest psdCa-O is determined and the putative sites within $3.5\text{Å}$ (Center-Center distance) from it are deleted. The procedure is repeated until the vector is empty. The computational complexity is dependent on the number of total comparisons, which is smaller than $O(n^2)$.
2.2.5 General algorithm analysis

The computational complexity of the algorithm is $O(n + n^2 + nm)$ from the three parts mentioned above, which can be simplified to be $O(nm)$ since it is the dominant component.

2.2.6 Performance measurement

A qualified clique is a true prediction (TP) if its CC falls into the cutoff distance (3.5 or 1.0Å in this study) from a documented calcium ion in the crystal structure. A documented calcium-binding site is a true predicted site (TPS) if there is any prediction within the cutoff distance from this site. The performance of the method is evaluated by Site Sensitivity (SEN), Site Selectivity (SEL), and Redundancy (RE), which represent the percentage of TPS in the total sites, the percentage of TP in the total predictions (hits), and the true predictions per site, respectively. The value of RE is not less than 1.

\[
SEN = \frac{TPS}{TotalSites} \quad (2.1)
\]
\[
SEL = \frac{TP}{TotalHits} \quad (2.2)
\]
\[
RE = \frac{TP}{TPS} \quad (2.3)
\]

2.3 Results and Discussions

<table>
<thead>
<tr>
<th>Dataset</th>
<th>Total proteins</th>
<th>Total sites</th>
<th>Proteins (multiple sites)</th>
<th>$CN \leq 4$</th>
<th>$CN = 3$</th>
<th>$CN \leq 2$</th>
</tr>
</thead>
<tbody>
<tr>
<td>I</td>
<td>32</td>
<td>62</td>
<td>18</td>
<td>55</td>
<td>7</td>
<td>0</td>
</tr>
<tr>
<td>II</td>
<td>54</td>
<td>91</td>
<td>26</td>
<td>80</td>
<td>8</td>
<td>3</td>
</tr>
<tr>
<td>III</td>
<td>44</td>
<td>94</td>
<td>27</td>
<td>81</td>
<td>7</td>
<td>6</td>
</tr>
</tbody>
</table>

Analysis of calcium-binding sites in three datasets. As shown in Tab. 2.3 there are a total of 123 proteins and 231 calcium-binding sites in three datasets. There are three proteins (1OVA, 2POR, and 4SBV) in both Datasets I and II and four proteins (1CEL, 1ESL, 1KIT, and 1SRA) in both
Datasets II and III [61, 75, 77]. These three datasets represent different classes of calcium-binding sites with different protein fold topologies (Fig. 2.3). For example, continuous sites as in calmodulin (3CLN) with four EF-hand motifs are largely helical [4] and lectin (2TEP) is predominately -sheet [79]. Semi-continuous sites such as in galactose-binding protein (1GCG) [33] or penicillin acylase (1AI4) [25] and discontinuous sites such as in annexin (1ALA) [10] or cellulase (1CEL) [24] are also included. About 55% of proteins in the three datasets contain two or more calcium-binding sites. Some proteins, such as calmodulin, do not have calcium-binding sites with shared ligand residues although the calcium binding process of this protein is tightly cooperative [4]. In other proteins, the clustered calcium ions share ligand residues and even ligand oxygen atoms. For example, mannose-binding protein A (2MSB), [105] thermolysin (1TMN, 1HYT and 8TLN), [42, 45, 74] neutral protease (1NPC), [90] serum amyloid P component (1SAC), [29] and carboxypeptidase T (1OBR) [95] in the datasets have calcium ions that share the same oxygen atoms Fig. 2.3. It is important that the datasets contain proteins representing different protein folds and families and contain different types of calcium-binding sites. For example, the data set III from Pidcock and Moore only contains one EF-hand protein (1SRA) [44] to overcome the bias of EF-hand proteins in the protein data bank [77].

As shown in Tab. 2.3, 55 (89%) of 62 calcium-binding sites in 32 proteins in Dataset I have 4 or more oxygen ligands within 3.5 Å of calcium, including 5 (16%) EF-hand proteins. This dataset was used for parameter adjustment. Dataset II contains 54 proteins with 91 sites, 80 (88%) of which have 4 or more oxygen ligands including 2 (4%) EF-hand proteins. Dataset III from Pidcock and Moore contains 94 sites in 44 proteins representing all classes and folds of calcium-binding proteins, 81 (86%) of which have four or more oxygen ligands.

2.3.1 Parameter optimization

Oxygen clusters are used to identify the potential calcium-binding sites in the GG program. The oxygen atoms from water are not included since 1) the inclusion of water molecules results in tremendous false predictions in the bulk solution and 2) the NMR and modeling structures typically do not contain the water molecules. However, the oxygen atoms from the cofactors, such as sugars or lipids, are included in this study. To achieve high accuracy and speed, clique Size, O-O Cutoff,
Figure 2.3: The calmodulin (A, 3CLN), galactose-binding protein (B, 1GCG), flavodoxin (C, 1AG9), and serum amyloid P component (D, 1SAC) represent different classes of calcium-binding sites including different ligand distributions (continuous, semi-continuous, and discontinuous), binding numbers (single site, independent multiple sites, and sites with shared ligands), cofactor conditions, and protein sizes. The green balls are calcium ions. The ligand oxygen atoms are from proteins (red balls) and cofactors (orange balls). The protein frames are in light blue while the cofactors are in orange. The pictures are generated by PyMol (DeLano Scientific).
PsdCa-O range, and D-filter in GG algorithm have been optimized using Dataset I.

Clique Size

In order to allow a clique in graphs to represent the calcium-binding location in a protein, the vertices and the edges should accurately describe the calcium-binding ligands and the relationships among them. To reveal key features of calcium-binding coordination and increase the speed of calculation, the model of calcium binding is simplified by using the minimal required cliques. As shown in Tab. 2.3, more than 85% of the sites contain 4 or more close oxygen ligands. The use of clique size 3 resulted in many false positive cases and the increase of computational complexity (data not shown) although the use of clique size \( \geq 4 \) resulted in many false negative cases in GG. To focus on the main features of calcium-binding proteins, we therefore chose the clique size 4 in this study. It is expected that the calcium-binding sites with 3 or fewer ligands cannot be found. However, we will show that part of these sites can still be identified. To better illustrate GG, the datasets were analyzed either including or excluding the sites that have 3 or fewer ligand oxygen atoms.

O-O Cutoff

The upper limit of O-O distance is restricted by the O-O cutoff while the lower limit is determined by van der Waals radius and thus is not specified in the algorithm. Theoretically, the upper limit of the O-O distance is no more than twice the maximum Ca-O distance. Most of the ligand oxygen atoms are in the distance of 2-3 Å to the calcium [77, 109] Fig. 2.4 shows the site sensitivity (SEN) of the GG algorithm using a series of O-O cutoff from 4 to 7 Å for Dataset I. When the cutoff is 5.0 Å, 89% (55/62) of the sites were identified. Among the 55 sites with 4 or more ligands, 52 (94%) have been identified within 3.5 Å and 51 (93%) within 1 Å to the documented ions. This result is consistent with the statistical study that shows the average Ca-O distance in calcium-binding sites is about 2.4 Å, [77, 109] suggesting that a cutoff of 5 Å covers most of the O-O distances. The longer O-O distances mainly originate from longer O-Ca distances. Further increase of the O-O cutoff results in the identification of more calcium-binding sites (within 3.5 Å to the real ions). At the cutoff of 6.0 Å, the SEN reaches 100% for Dataset I. Therefore, an O-O cutoff of 6.0 Å was
Figure 2.4: The site sensitivity (the true predicted sites in the total sites) as a function of O-O distance cutoff

used for all sites. 2.4.

_PsdCa-O range and D-filter_

To eliminate false positive predictions, the psdCa-O range and D-filter were introduced into the GG program. A clique is considered to be a potential calcium-binding site only when the psdCa-O is in a given range as long as there are no other atoms within a distance of D-filter to the circumcenter. SEN exhibits an upward trend while SEL shows a downward trend with the increase of psdCa-O range. The best performance was obtained when D-filter equals the psdCa-O of each clique. Corresponding to the average Ca-O distance of 2.4 Å, 48 sites (77% of all sites and 87% of the
sites with 4 or more ligands) in Dataset I are identified within 1 Å to the documented ions. If the psdCa-O of 2.3-2.5 Å is used, the corresponding SEL is 96%. When the psdCa-O range is enlarged to 2.1-2.7 Å, the SEN increases to 95% and the SEL is 87% for the sites with 4 or more ligands. In this case, three calcium-binding sites in ovalbumin (1OVA), (Stein, et al., 1991) staphylococcus nuclease (1SNC), [66] and substilisin BPN (2ST1) [12] are not identified. When the psdCa-O range is 1.8-3.0 Å, the only unidentified site is that in ovalbumin. Under such conditions, 4 of 7 sites with 3 or fewer ligands in the Dataset I have also been identified within 3.5 Å of the documented ions but none of them is within 1.0 Å. Taken together, using an O-O cutoff of 6.0 Å, a psdCa-O range of 1.8-3.0 Å and a D-filter equal to the psdCa-O, the optimal performance for the prediction of calcium-binding sites in Dataset I possesses a SEN of 95% and a SEL of 86% within 3.5 Å or a SEN of 87% and a SEL of 68% within 1.0 Å to the documented ions.

2.3.2 GG algorithm predicts calcium-binding sites with high SEN and SEL

Using an O-O cutoff of 6.0 Å, a psdCa-O range of 1.8-3.0 Å and a D-filter equal to psdCa-O, a SEN of 91% and a SEL of 77% have been achieved for the prediction of 91 calcium-binding sites in Dataset II when a true prediction is assumed if the predicted calcium is within 3.5 Å of the documented calcium ions. A SEN of 87% and a SEL of 74% have been achieved in Dataset III under the same conditions (Fig. 2.5A). If the standard of a true prediction is that the predicted calcium is within 1 Å of the documented calcium ions, the SEN and SEL are 84% and 59% for Dataset II and 80% and 55% for Dataset III, respectively (Fig. 2.5B). Similar to Dataset I, 6 of 11 sites with 3 or fewer ligands in Dataset II and 4 of 13 in Dataset III have been identified within 3.5 Å of the real calcium but none of the predictions is within 1 Å of the real calcium. On the other hand, if only considering the calcium-binding sites with 4 or more ligands as real sites, the SEN for Datasets I, II, and III increases to 98, 95, and 93%, respectively, with the true prediction standard of within 1.0 Å of the documented calcium ions (Fig. 2.5 ). The slightly different performances for the three datasets reflect the different distributions of calcium-binding sites (Tab. 2.3, where Dataset III has the highest populations of sites with 3 or fewer oxygen ligands and Dataset II contains 14 non-calcium-binding proteins [61, 75, 77]. Although the performance of the prediction was lower by including these less popular calcium-binding sites or non-calcium-binding proteins, it confirms
Figure 2.5: The sensitivity (black bars) and selectivity (grey bars) of the GG program with the O-O distance cutoff of 6.0 Å and the psdCa-O distance range cutoff of 1.8-3.0 Å. The standard of a true prediction is that the predicted calcium position is within 3.5 Å (A) or 1.0 Å (B) threshold distance to the documented calcium. The results of all sites are shown on the left and the results only considering the sites with 4 or more ligands are shown on the right.

the practicability of high site selectivity, sensitivity and accuracy of our GG algorithm. 2.5. The redundancy of predicted cliques (RE) is about 14 for all three datasets. This value agrees well with
the average coordination numbers of 6 for calcium in proteins. Theoretically, a clique of size 6 has 15 sub-cliques of size 4. The predictions of a single site are developed by combining sub-cliques into one using the merging algorithm (Method). The final RE values after the merging are 1.15, 1.18, and 1.21 for the Datasets I, II, and III, respectively. This low RE together with the high SEN and SEL facilitates the subsequent analysis and is advantageous in predicting large datasets. GG is very effective for sites with 4 or more oxygen ligand atoms. After using the merging algorithm, 54 of 55 sites in Dataset I, 78 of 80 sites in Dataset II, and 78 of 81 sites in Dataset III with 4 or more ligands have been uniquely identified.

Calcium-binding sites that were not identified by our GG algorithm usually possess abnormal Ca-O distances. 1OVA in Datasets I and II, 1CEL in Datasets II and III, 1DJX and 1SCM in Dataset II, and 1BJR and 1AG9 in Dataset III were not identified. 1OVA has a long Ca-O distance of 3.3 Å between the calcium and the O1 of attached phosphate at S350 as well as two very short distances (1.6 Å from phosphate O3 at S350 and 1.9 Å from O 2 of E201) [91]. In 1CEL, four ligands are from two glutamates (E325 and E295) while the C atoms of both residues are closer to the calcium ion than O 1 of E325 [24]. In 1DJX, the C is closer to the calcium than the ligand O 2 of D653 [30]. In 1SCM, there is a carbon atom (C of D28) located at 2.8 Å to the Ca502 [106]. These sites are rejected by the psdCa-O range and D-filter. In 1BJR, the O-O distance between the two ligands of Ca290, the backbones of S15 and A273, is 6.02 Å, exceed the O-O cutoff of 6.0 Å [88]. In 1AG9, the closest oxygen from the protein to Ca350 is 4.4 Å away. (Hoover and Ludwig, 1997) The calcium is chelated by four oxygen atoms from the cofactor BTB (Bis-2-Hydroxy-Imino-Tris-Hydroxymethyl-Methane), which are located at one side of the calcium ion [48] Thus, the clique center is far from the calcium ions and is either rejected or not counted as a true prediction. To maintain the sensitivity, the O-O cutoff and the psdCa-O range have to be increased. As a result, the CC shifts away from the calcium position, the SEL decreases, and the RE value increases.

2.3.3 The prediction of calcium-binding sites with three or less ligand oxygen atoms

As shown in Tab. 2.3, 13% of 231 calcium-binding sites contain 3 or fewer oxygen atoms. Oxygen atoms from water were excluded for the prediction due to their limited contribution to calcium-binding affinity [35, 72, 75]. Since we use clique size of 4, it is expected that the calcium-binding
sites with 3 or fewer ligands cannot be found. Surprisingly, 4 of 7 (57%), 6 of 11 (54%), and 4 of 13 (30%) of these sites in Dataset I, II, III, respectively, have been identified although it is relatively less sensitive. The results suggest that other non-binding oxygen atoms are crowded around the calcium-binding sites in addition to the ligand oxygen atoms. Although these oxygen atoms do not chelate calcium directly, they can form cliques together with the ligand oxygen atoms and result in a partial identification of less-coordinated sites using a clique size of 4.

2.3.4 Performance Comparison

Several methods such as Feature and SeqFeature based on the statistic function of the physiochemical environment around functional sites have been developed for predicting calcium-binding sites [5, 52, 80, 89, 104]. These methods predict the calcium-binding sites by a scoring system taking into account multiple properties such as secondary structures, chemical groups, and atom types for which high sensitivity and selectivity have been claimed. The same dataset investigated by Altman et al. with Feature was analyzed by GG program. The standard of true prediction of calcium-binding sites is within 6.0 Å of the documented calcium for FEATURE [61] in contrast to 1.0 Å for our prediction using GG. Even though, the SEN from FEATURE is similar to that from the GG with the sacrifice of SEL [61]. For the sites with 4 or more oxygen ligand atoms, we have achieved SEN of 95% within 1 Å of the structural resolution. Our results have shown that the GG program not only provided high prediction sensitivity without the use of vast statistical properties (i.e. based on oxygen cluster only) but also retain a greater site resolution to the documented calcium positions. This is very important for the application of testing the function of the proteins with required resolution and accuracy.

Using a somewhat less diversified dataset, the best geometry-based prediction of calcium-binding sites at present has been achieved using the valence study by Nayal and Di Cera [75]. In their study, points with high valences in a fine grid are output as potential calcium positions. A SEN of 93% and a SEL of 95% have been reported when the standard of true prediction is within 3.5 Å to the documented calcium. This SEN is similar to that from GG (95%) while the SEL is higher than that from GG (86%). There are several major improvements of our GG algorithm. First, the SEL reported by Nayal and Cera has definitions for selectivity different from GG. In GG, the SEL
is defined as the total true predictions among the total predictions. When one site has multiple predictions, the selectivity of the sites is different from the selectivity of the predictions. After merging the related predictions, the site selectivity of the GG program is 40% for the Dataset I. In the valence study by Nayal and Cera, the SEL is derived from the predicted points [75]. It is expected that the qualified points in a calcium-binding pocket are much denser than that out of the pockets. As indicated by RE, the SEL might be significantly affected after merging the points in their study. Second, the RE and the total predictions of GG are much smaller than that in the valence study. The RE from the valence study is about 69, almost 5 times of that from GG. The total predictions are 4018 in the valence study and 808 in GG [75]. In one aspect, the high RE indicates the high deviation of the SEL. More importantly, if further analysis is required, it will suffer from the large result data file. The small result file together with only 20% of false positives from the GG algorithm will be a beneficial and practical for the further evaluation. Third, in the valence study, a total of 7.2 million points have been calculated while only 51258 points in the GG are calculated [75]. Correspondingly, the GG program is almost 600-time faster than the valence calculation. For example, it took 141 CPU seconds for the valence calculation program (vale.exe) to examine mannose-binding protein A (2MSB) using an Intel Celeron M 1.30 GHz processor, which prevents the practical use of this algorithm to analyze large databases. In contrast, GG only requires 0.24 CPU seconds for the same analysis. This high speed is important for the application to the prediction of calcium-binding sites using vast structural genomic information. Fourth, the GG program is very sensitive to sites with 4 or more close ligands. Even with a higher standard of 1.0 Å, more than 93% of the sites have been identified in all datasets. Since the majority of the natural calcium sites have 4 or more ligands and the remaining either require the cofactors or bind weakly, the results indicate that the GG program is a powerful tool for most of the calcium-binding proteins. Moreover, our GG algorithm can still predict calcium-binding sites with 3 or less oxygen ligand atoms with regular Ca-O distances.

2.3.5 A cluster of four or more oxygen atoms has a high potential for calcium binding

The three datasets analyzed in this study contain a total of 231 calcium-binding sites in 123 proteins. They cover all major classes of the natural calcium-binding proteins and different protein frames
as reported by Pidcock and Moore [77]. These calcium-binding sites also cover a broad range of calcium affinity with dissociate constants from 10-10 (e.g. subtilisin, 2ST1) [48, 64] to 10-4 M (e.g. thermolysin, 8TLN) [42, 45, 64, 74] in addition to 14 non-calcium-binding proteins in Dataset II. As shown in Tab. 2.3, more than 85% of the sites contain 4 or more ligands from proteins and cofactors. Less than 15% of the sites contain 3 and fewer ligand oxygen atoms. A detailed examination revealed that part of these sites is really just disordered higher coordination sites, which can be identified using 4-atom cliques with increased O-O cutoff distances. For example, 14 of 31 sites with 3 or fewer oxygen ligand atoms were identified within 3.5 Å of the documented calcium ions. Since these sites are normally located at the protein surfaces, the development of a program based on GG to predict them is in progress by combining the 3-atom cliques with the calcium-specific valence calculation and adding a structured water molecule. We have noticed that the sidechain atoms in NMR structures usually possess less defined coordinations than the backbones due to lack of structural constraints. To overcome such uncertainties, we plan to further integrate the sidechain re-packing programs with the GG program to allow the visualization of calcium-binding sites in structures with lower local resolutions [77, 109]. The chelate of calcium in proteins is largely of an electrostatic nature, which is a balance of Ca-O charge-charge attraction and ligand O-O repulsion in addition to the constraint from the sidechain configurations and protein frames [69, 78]. Efforts have been made previously to locate the calcium-binding sites using an energy-based united residue potential calculation [55]. Since ligand oxygen atoms could be from highly charged carboxyl and partially charged carbonyl and hydroxyl groups in addition to water and cofactors, the arrangement of 4 or more ligand atoms can be very different from traditional polygon coordination of metal ions such as zinc, magnesium, or iron. The high sensitivity and selectivity of the GG program suggest that a cluster of four or more oxygen atoms is not only necessary but also sufficient for most calcium binding. The results reveal that clusters of oxygen atoms without other atoms inside can be classified to identify calcium-binding sites. The metal-binding sites using non-oxygen ligands will not be identified based on the oxygen clusters. For example, copper and manganese also bind to nitrogen atoms in addition to oxygen atoms. To distinguish the calcium-binding sites from other all-oxygen-ligand metal-binding sites requires further consideration of other factors such as the charge nature, the knowledge of hydrogen waters, the size of different metal ions, and the binding geometry. For example, magnesium possesses a relatively smaller ionic radius (Mg 0.65 Å vs. Ca
0.99 Å), a shorter bond distance (Mg-O 2.1 Å vs. Ca-O 2.4 Å), and less protein ligands due to the hydration nature of the magnesium ions.

The previous version of the GG program predicts the calcium-binding sites without a significant movement of oxygen atoms or global conformational changes. It is capable of being incorporated with other methods or altered parameters for better performance. It can be a powerful tool for automated analysis of large structural genomic databases with sensitivity greater than 85% in addition to its capability of providing structural resolution (within 1 Å of the documented ions and with high speed). More importantly, our developed algorithm can be directly applied to predict calcium-binding sites in proteins and receptors with weak calcium affinities overcoming the major limitation of the current methodology. To reveal the calcium binding in the NMR structure of our designed calcium-binding protein Ca.CD2, we have previously applied the same concept of oxygen cluster to facilitate the visualization of the metal coordination. The manganese-relaxation NMR experiments clearly validated this prediction strategy [108]. The development of the GG program will significantly accelerate such processes of locating calcium ions in the NMR structure. Furthermore, our study using graph theory has revealed new features of calcium binding in proteins and the developed program may facilitate the understanding and prediction of functions of calcium roles in biological systems based on oxygen coordination, which can be provided by homology modeling based on the sequence information. The GG program is available at http://chemistry.gsu.edu/faculty/Yang/GG.htm or http://mathstat.gsu.edu/magtcc/research/06/Calcium/GG.html.
Chapter 3

IDENTIFYING CALCIUM-BINDING SITES WITH GEOMETRIC AND CHEMICAL CRITERIA AT OXYGEN-CARBON SHELL

In the Chapter 2, we describe a graph theory and geometry approach to improve the accuracy for predicting calcium-binding sites based on the valence function. However, in order to explore in-depth the geometric properties of the oxygen-carbon shells of calcium-binding sites, we enhance our previous approach at a high level to find biggest local oxygen clusters with a graph algorithm to find maximal cliques and propose a new geometric criterion embedding the bidentate property to filter non calcium-binding oxygen clusters. In addition, we apply some motifs of residue combinations as another filter to exclude non calcium-binding oxygen clusters possibly formed by hydrogen bonds and obtain higher site selectivity without trading off site sensitivity. The experiments demonstrate good predictive performance on both old and new datasets.

3.1 Introduction

Yamashita et al. developed an algorithm based on the hydrophobicity/hydrophilicity contrast function and successfully predicted ion-binding sites in a number of cases [107]. Nayal and Di Cera improved the result by establishing a new simple function, valence function. Their approach gives fair accurate on identifying many points on the space may be potential sites [75]. Consequently, one calcium biding site may correspond to many potential predictions. Unfortunately, there is lack of way to filter out non-calcium-binding sites as the Nayal and Di Cera pointed out that the points where the valence function reaches maximal values may not be closer to calcium binding sites than the points which have smaller valence value. Altman et al. also developed a statistical method to identify the microenvironment of metal binding sites, but their prediction standard is 6.0Å from the real calcium, which is too loose to position the calcium cation [5, 61, 60, 104].

To overcome the shortage of the above approaches, we need to understand some geometric properties as well as other properties of calcium binding sites. However, as mentioned by the literature
references [107], it is impossible to characterize the crystal geometric properties of oxygen atoms surrounding calcium binding sites in the traditional way. Despite the difficulty of characterizing the geometric properties of all oxygen atoms surrounding a calcium binding sites, We discovered there are always four oxygen atoms on the sphere with certain size centered at the calcium binding sites [23, 22]. Based on this discovery, they established a graph algorithm to identify all clusters with four oxygen atoms and then use a geometric algorithm to filter out the non calcium binding sites and identify the calcium binding sites if the four oxygen atoms satisfy the condition.

Schymkowitz et al. developed a broad method to predict water and metal binding sites as well as their binding affinities from the Fold-X field [84]. They obtained a high accuracy of 97% for the prediction with the precision on a certain dataset of 115 proteins with 244 calcium-binding sites [84]. However, their calculation is somewhat complicated in both fusion and optimization phases [84].

We developed a graph theory and geometry (GG) approach for rapidly identifying calcium-binding sites in proteins, which is detailed in the Chapter 2. The previous version is called GG1.0. The current version of the GG approach, called GG2.0, employs the maximal clique algorithm to find biggest local oxygen clusters/cliques and uses an optimization tool to calculate the geometric filter related to the ratio between the size of the first shell and the second shell of calcium-binding sites. In this paper, we include the geometric properties of carbon shell as well as oxygen shell and analyze these properties with a special geometric angle in calcium-binding sites. In order to do so, we find all maximal clusters with at least 4 oxygen atoms from proteins. We also establish new data sets to conform to the requirements as the Fold-X method.

In addition, we observe that some oxygen clusters satisfying geometric criteria are not calcium binding sites with some fixed residue combinations. A site sensitivity of 98% with a site selectivity of 86% is obtained on the newly-created training set with adjusted parameters on geometric properties. The average deviation of predicted calcium locations is 0.66 Å. To validate the performance of GG2.0, we apply it on the test set and achieve a site sensitivity of 100% and a site selectivity of 89%. More importantly, the GG2.0 approach reveals the geometric property of the first-second shell of calcium-binding sites and some motifs of residues consisting of non calcium-binding clusters.
3.2 Methods

3.2.1 Datasets

To acquire a high resolution and non-homology dataset of proteins with calcium-binding sites, we query from the metalloprotein database and browser (MDB) [17] with the conditions that every calcium-binding protein structure from PDB [9] has a resolution less than 2.0 Å from X-ray crystallography, each site has a coordination number greater than three excluding water oxygen, and the PDB entry must be in the PDBSELECT non-homology list (Hobohm and Sander). The retrieved dataset contains 163 proteins of 345 sites. Because the current state-of-the-art method, Fold-X, targets at the calcium-binding sites with at least four coordinated atom number [84], we exclude the PDB entries containing a calcium-binding site which does not conform to the same requirement as the Fold-X method. Finally, the training dataset contains 121 protein structure files with all 240 calcium-binding sites. The test dataset contains 20 proteins listed in the literature of Fold-X method in 2005 [84]. In this test dataset, there are four calcium binding sites which have the coordinated ligand number less than four each and are not taken into account for calculating prediction accuracy.

3.2.2 Graph algorithm

For a given protein structure as shown in Fig. 3.1, the coordinates in the PDB file of oxygen atoms including the hetero atoms but excluding those from water molecules were extracted first for the analysis. The distances between every two oxygen atoms are calculated. A graph, $G(V,E)$, is constructed accordingly, in which each vertex in $V$ represents an oxygen atom or a carbon atom which has a covalent bond with an oxygen atom and each edge in $E$ represents either a covalent bond between an oxygen atom and a carbon atom or a relation between a pair of oxygen atoms apart within an O-O cutoff distance. Only the edges among oxygen atoms are weighted by the Euclidean distance. The graph construction time is $O(n^2)$, where $n = |V(G)|$ is the number of oxygen atoms. A clique $Q$ is maximal if there is no clique containing $Q$ as a proper subset.

The redundant cliques will lower the efficiency of the program and cost more efforts in analyzing the results. In principal, a clique with a size of 5 and 6 contains 5 and 15 sub-cliques of a size of 4, respectively. Therefore, local maximal cliques instead of cliques with fixed sizes which were
Figure 3.1: The schematic model of GG2.0 Program is shown. The positions of oxygen atoms in green (non-water oxygen) and orange (water oxygen) and bonded carbon atoms in blue are extracted from the protein structure. An edge is given between two oxygen atoms if the distance between them is less than the cutoff distance. A putative calcium-binding site is an oxygen cluster/clique with a constraint on the average radius of two clusters (right bottom).
once employed in the GG1.0 [22] are searched in this study. To search for a clique with size $N$ or above, all oxygen atoms possess less than $N-1$ edges have been eliminated first. The local maximal cliques are identified in the remaining atoms following the procedure shown in Scheme 1 which was developed by Bron and Kerbosch [21]. We choose the simple and classical clique-finding algorithm because the time complexity is bounded by linear time for this special graph constructed. The algorithm combines a backtracking search with a branch and bound technique. The backtracking search is to find the maximal cliques recursively. The branch and bound technique is used to stop some useless search failing to find a new maximal clique in advance. There are three important sets in this algorithm:

1. Candidate Clique is a set of vertices connecting each other. The set can be either added or reduced by a vertex along the backtracking tree.

2. Candidates is a set of vertices qualified for adding to the set of Candidate Clique.

3. Old Candidates is the set of vertices which were added to Candidate Clique before and not in it now. The algorithm executes recursively addition on Candidate clique with one of vertices from Candidates. In the mean time it keeps the vertices connecting to the candidate vertex in Candidates and Old Candidates. When Candidates is empty, Candidate Clique can not be extended any more and is satisfied for the maximal clique. The early stop condition is reached once a vertex in Old Candidates connects all vertices of Candidates. It is foreseen that maximal cliques can not be found in this branch, which can be proved by counter-proof. Assuming there is a maximal clique founded in the branch, a vertex can be added to the maximal clique to make a bigger clique. This violates the definition of maximal cliques.

Let $\text{Cost}(G)$ be the cost of the computational complexity of this algorithm. As a direct consequence of lines 2, 10, 12, 13, and 18 of the pseudo code (Fig. 3.2), we have

$$\text{Cost}(G) \leq O(n\Delta^2)$$

where $n = |V(G)|$ is the number of oxygen atoms and $\Delta$ is the maximum degree in $G$. For a general graph, $\Delta$ can be as big as $n-1$ so that the computational complexity is $O(n^n)$. However, we prove that $\Delta$ is bounded above by a constant for the graphs created from PDB file. Then, from
1. **Function** `findMaxClique (CandidateClique, Candidates, OldCandidates)`

2. **Loop** through every vertex \( l \) in \( \text{OldCandidates} \)
3. **Loop** through every vertex \( m \) in \( \text{Candidates} \)
4. If there is no edge between \( l \) and \( m \)
   Flag := false; **break**;
5. If Flag equals true
   **return**; //branch and bound
6. If Candidates is empty //base case for recursive function
7. CandidateClique is a maximal clique;
8. **Loop** through every vertex \( k \) in \( \text{Candidates} \)
9. CandidateClique := \( k \);
10. **Loop** through every vertex \( o \) other than \( k \) in \( \text{Candidates} \)
11. If there is an edge in the pair \( (k, o) \)
12. newCandidates := \( o \);
13. **Loop** through every vertex \( p \) other than \( k \) in \( \text{OldCandidates} \)
14. If there is an edge in the pair \( (k, p) \)
15. newOldCandidates := \( p \);
16. `findMaxClique (CandidateClique, newCandidates, newOldCandidates);`
17. remove \( k \) from CandidateClique;
18. `OldCandidate := k;`

Figure 3.2: The procedure of finding the maximal cliques.
the above inequality, \( \text{Cost}(G) \leq O(n) \), which implies that the job to find cliques can be finished in linear time. Here we restrict four as the minimum size of maximal cliques because every target binding site has at least four protein oxygen ligands.

### 3.2.3 A geometric criterion

After oxygen clusters/cliques’ finding, we can obtain a carbon cluster around every oxygen cluster because each oxygen connects one carbon atom. Each oxygen cluster could have a geometric point called LVP from which the distances to every other atom/vertex of the cluster/clique have a smallest variance. For every oxygen cluster, there is a corresponding carbon cluster surrounding it. These two clusters are thus called twin clusters and the LVP of the oxygen cluster is chosen as the center of the twin clusters. In fact, we also calculate the LVP of a carbon cluster to analyze calcium binding site and results (data not shown) are not as good as the above-mentioned selection of LVP. We use the optimization function of fminsearch in the software of Matlab7.0 to obtain the coordinates of the LVP of a cluster. A radius of oxygen/carbon (RO/RC) can be calculated as follows.

\[
RO = \frac{\sum_{i=1}^{k} \text{dist}(LVP, O)}{k} \tag{3.1}
\]
\[
RC = \frac{\sum_{i=1}^{k} \text{dist}(LVP, C)}{k} \tag{3.2}
\]

where \( \text{dist}(LVP, O) \) \( \text{dist}(LVP, C) \) labels the distance between the LVP and each oxygen/carbon ligand, and \( k \) is the number of vertices of a cluster. The value of \( \frac{RO}{RC} \) reflects the size of an oxygen/carbon shell to some extent. There is a ratio between the RO and the RC for every twin clusters, briefed as \( r_{RO_RC} \). To eliminate false positives, we use \( r_{RO_RC} \) as a filter within some range for a putative calcium-binding site. From the experiments, the results (data not shown) are not as good as those using the adjusted \( r_{RO_RC} \), \( ar_{RO_RC} \), because the carbon shell will become smaller when a calcium-binding site has a bidentate residue as ligand and the \( ar_{RO_RC} \) is bigger. In this way, the \( ar_{RO_RC} \) is chosen to replace the \( r_{RO_RC} \) as the filter. The \( r_{RO_RC} \) and \( ar_{RO_RC} \) are calculated as follows.

\[
r_{RO_RC} = \frac{RO}{RC} \tag{3.3}
\]
\[
ar_{RO_RC} = r_{RO_RC} - 0.5NB \tag{3.4}
\]
where NB shows the number of bidentate residue(s) in a putative calcium-binding site. The complexity of this optimization algorithm is $O(n)$ because only constant calculations are needed for every pair of clusters and there are $O(n)$ pairs of clusters in total.

3.2.4 A chemical criterion

From the experiment results, some oxygen clusters satisfying the geometric criteria are not around calcium binding sites. In those clusters with size four, we observe some patterns of residue combination for non calcium binding sites. The patterns are summarized into two rules as the chemical criteria as follows.

1. If a cluster contains a backbone carbonyl oxygen atom, it is considered to be putative sites.

2. If a cluster contains two side-chain carboxylate atoms from different residues, the cluster is considered to be putative sites. The complexity of this step is also $O(n)$ because only constant calculations are needed for every twin clusters and there are $O(n)$ pairs of clusters in total.

3.2.5 Removal of the redundant predictions

To remove redundant predictions in one location, we develop a merging algorithm. All putative sites in a protein are recorded as a vector and sorted by clique size and RO in order. The one with the largest cluster/clique is determined and other clusters/cliques sharing two or more common vertices are deleted. If there are two clusters with same size, we choose the one with smaller RO. The procedure is repeated until the vector is empty. The computational complexity is dependent on the number of total comparisons, which is $O(n \log n)$.

3.2.6 General algorithm analysis

The computational complexity of the algorithm is $O(n^2 + n + n \log n)$ from the four parts mentioned above, which can be simplified to be $O(n^2)$ since it is the dominant component.
3.2.7 Performance measurement

A qualified clique is a true prediction (TP) if its LVP falls into the cutoff distance (2.0Å in this study) from a documented calcium ion in the crystal structure. A documented calcium-binding site is a true predicted site (TPS) if there is any prediction within the cutoff distance from this site. The performance of the method is evaluated by Site Sensitivity (SEN), Site Selectivity (SEL), and Deviation (Dv), which represent the percentage of TPS in the total sites, the percentage of TP in the total predictions (hits), and the average distance between predicted location and documented location, respectively.

\[
Dv = \frac{D_{vs}}{Total\, Hits}
\]  

(3.5)

3.3 Results and Discussions

3.3.1 Parameter setting

Clique Size

Similarly to the GG1.0, we use four as the threshold of clique size, which means only clusters with four or more oxygens are considered as putative calcium-binding sites. Because only a small portion of calcium-binding sites with less than four ligands, we aim at calcium-binding sites with four or more ligands.

O-O Cutoff

From the statistical analysis of the GG1.0, we use 6.0 Å as the O-O cutoff since we take the oxygen shell as 3.0Å from the calcium cation and the maximum distance between two oxygen ligands in the oxygen shell is 6.0 Å.

ar_RO_RC

As shown in Fig. 3.3a, the plot shows the histogram of RO and RC on the new dataset. It indicates the distribution of RO and RC on the new dataset similar to the analysis from Nayal and Di Cera [75] on the old dataset. The peak value for RO is within the range of 2.4 – 2.5Å while the peak value
Figure 3.3: The histogram of RO and RC (a), the histogram of bidentate r_RO_RC, bidentate r_RO_RC and r_RO_RC (b), and the histogram of bidentate ar_RO_RC, bidentate ar_RO_RC and ar_RO_RC (c).
for RC is within the range of 3.2 – 3.3 Å. In a calcium-binding site, the RO is always smaller than the RC because the former represents the size of the oxygen shell whereas the latter represents the size of the carbon shell. Therefore, $r_{RO \text{ RC}}$ is less than 1. The Fig. 3.3b indicates that $r_{RO \text{ RC}}$ has the range between 62% and 82%. After the adjustment of $r_{RO \text{ RC}}$ according to the bidendate property of oxygen shell, $ar_{RO \text{ RC}}$ falls into the smaller range between 62% and 74% with only one exception that the $ar_{RO \text{ RC}}$ is 56% in the calcium binding site with three bidentate ligand residue as seen in the Figure 2c. We choose a series of $ar_{RO \text{ RC}}$ values from 71% to 75% as the threshold value of the geometric filter, which means if $ar_{RO \text{ RC}}$ is greater than the threshold value, the oxygen cluster is not considered as a putative calcium-binding site; otherwise it is. As shown in Figure 3, with the increase of the $ar_{RO \text{ RC}}$, the SEN increases but the SEL decreases. When it reaches 75%, the SEN does not increase but the SEL decreases. Therefore we adopt the value 74% as the threshold of the filter to exclude non calcium-binding oxygen clusters.

3.3.2 performance

As seen from the Fig. 3.4, without using the chemical filter, the GG2.0 can obtain the prediction accuracy of the SENs ranged from 92% to 98% with the SELs ranged from 87% to 78%. There is a trade-off between the SEN and the corresponding SEL. Because we prefer SEN with a higher value than SEL, so 74% is taken as the empirical value of $ar_{RO \text{ RC}}$ for the threshold of the filter. Although the SEN also reaches 98% at the $ar_{RO \text{ RC}}$ value of 75%, but the SEL decrease relative...
to the SEL at the $ar_{RO_{RC}}$ value of 75%. To explain how the geometric filter comes out, we analyze the correlation between the $r_{RO_{RC}}$ and the average angle of Ca-O-C in each calcium-binding site. As seen in Fig. 3.5, the angle decreases as the $r_{RO_{RC}}$ increases. In order to verify the negative linear relationship of these two parameters, we calculate the Pearson coefficient and obtain -0.97 which proves this strong correlation. Additionally, the distribution of bidentate and non-bidentate angles shows almost non-overlap range which matches the distribution of $r_{RO_{RC}}$ as seen in Fig. 4.10. In the meantime, it indirectly proves that $r_{RO_{RC}}$ become larger when the bidentate residue(s) join the calcium-binding site from the average angle of Ca-O-C. As a matter of fact, the $r_{RO_{RC}}$ can be calculated by solving the Equations 4.3 and 3.6.

\[
\text{dist}(C,O)^2 = \text{dist}(C,O)^2 + \text{dist}(C,O)^2 - 2 \times \text{dist}(C,O) \times \cos \text{Ang}_{ca-o-c} \quad (3.6)
\]

where dist(C,O) means the distance between the carbon and the oxygen atoms, dist(Ca,O) means the distance between the calcium cation and the oxygen atom, and dist(Ca,C) means the distance between the calcium cation and the carbon atom. To roughly calculate the $r_{RO_{RC}}$, we could take RO as the value of dist(Ca,O), RC as the value of dist(Ca, C), and the empirical value of 1.2 Å as the value of dist(C,O). It seems that a conflict exists between the statistical analysis and mathematical equations on the relationship between $r_{RO_{RC}}$ and angle Ca-O-C. In fact, dist(Ca, O) also changes with the change of angle Ca-O-C (data not shown), which may partly explain the quadratic equation (8). Using the chemical filter with the ar$_{RO_{RC}}$ of 75%, the GG2.0 still obtains the best site sensitivity of 98% while increases the SEL from 82% to 86%. The result is comparable to the result of the current of state-of-the-art method, Fold-X. This means that the chemical filter is the absolute filter to exclude non calcium-binding oxygen clusters. It is noted that ASP and GLU are from the charged residue group, and SER, THR, TYR, HIS, ASN and GLN are from the polar residue group, according to the classification of chemic properties on twenty basic amino acids. It is already known that hydrogen donor protein oxygen atoms are SER OG, TYR OH and THR OG1, and acceptor oxygen atoms are carboxyl oxygen of the side chain, ASN OD1, GLN OE1, ASP OD1, GLU OE1, SER OG, ASP OD2, GLU OE2, THR OG1, and TYR OH [96]. In particular, approximately 90% of the ‘bifurcated’ bonds are of ST/DE type [96]. "The term ‘bifurcated hydrogen bonds’ implies that hydrogen of the ‘rotating’ hydroxyl of Ser or Thr may interact with the two oxygen atoms of a carboxyl group” [96].
oxygen atoms may form a cluster within a O-O cutoff of 6.0 Å because of two hydrogen bonds, which provide strong supports for the chemical filter on some oxygen clusters with size four. After using the merge algorithm, we can finally annotate the calcium binding position with a Dv of 0.66 Å for each identified calcium binding site. Although somewhat lower than the result of Fold-X method [84], it is higher than all other previous methods such as the FEATURE [61] and the Valence methods [75]. In addition, our filters have a high selectivity. To test the reliability of the GG2.0, we apply it on the testing dataset. 48 out of 52 total calcium binding sites are identified with a high selectivity of 89%. It is worth pointing out that the site sensitivity would become 100% if we do not count the four calcium binding sites with protein ligand number less than four. From the supporting materials of the Fold-X method, it is hardly seen that Schymkowitz et al. count these calcium-binding sites as prediction accuracy (which is an arguable point) [84].
Chapter 4

FEATURE ANALYSIS ON CALCIUM-BINDING SITES

4.1 Introduction

The biological roles of proteins are inextricably linked to their three-dimensional tertiary structures, which in many cases, are achieved as a result of binding interactions with metal ions. In fact, nearly 40% of all proteins bind metals [35, 46, 49, 65, 87, 94], and a search for "metal binding proteins" in PubMed (http://www.ncbi.nlm.nih.gov/) returns over 87,000 entries.

For naturally-occurring metalloproteins, the binding of $Mg^{2+}$, $Zn^{2+}$ and $Ca^{2+}$ ions has been studied extensively with respect to binding geometry, structure, coordination number, and ligand preference [2, 3, 18, 19, 27, 26, 35, 38, 39, 40, 76, 99, 100]. Studies of peptide fragments have revealed further data regarding the binding affinities of $Ca^{2+}$-binding structures [69]. Despite significant variations in these values with respect to each metal ion, the majority of metal-binding sites are characterized by a central shell of hydrophilic ligands to chelate the ion, with a surrounding shell of hydrophobic residues [5, 107].

Ionic calcium is demonstrably one of the more relevant metals associated with biological systems. In the human body, more than 90% of the extant calcium can be found in bones and dental enamel as hydroxyapatite, and $CaCO_3$ is prevalent in the biomineralization of shells and corals [70]. Calcium ions bind to various $Ca^{2+}$-binding proteins (CaBP’s), which results in conformational changes, thus inducing function. These functions include muscle contraction, neurotransmitter release, enzyme activation and blood-clotting [43, 47, 68]. Calcium also participates in various functions related to the cell life cycle including both cell division and apoptosis, and its prevalence in both extracellular and intracellular fluids, where fluctuations of $Ca^{2+}$ concentrations in the latter, which normally range from sub-micromolar to millimolar levels, allows it to act as a secondary, or intracellular, messenger [16, 70, 76, 82].

The most common motif associated with $Ca^{2+}$-binding, the EF-Hand motif, was first described
by Kretsinger in 1973 [56], and is observed in over 50% of all $Ca^{2+}$-binding proteins. This motif is characterized by a helix-loop-helix structure comprised of approximately 30 amino acids, and this motif can be sub-divided into canonical EF-Hand (e.g. - calmodulin) and pseudo EF-Hands found in the S100 protein N-termini [112]. A thorough summary of the classification and evolution of EF-Hand proteins was presented by Kawasaki, Nakayama and Kretsinger [54].

The results of these studies have provided several key characteristics associated with $Ca^{2+}$-binding. Calcium ions, classified as hard Lewis acids, form ionic bonds predominantly with oxygen ligands, and the $Ca^{2+}$-O bond length ranges from 2.01-3.15 Å, with a mean value of $2.4 \pm 0.2\ Å$ [27, 35, 43, 93, 102]. Observed structures frequently exhibit irregular pentagonal-bipyrimid geometry (Figure 1), with 3 monodentate ligands and 1 ligand providing the 5 planar oxygen ligands. The $Ca^{2+}$ ion resides $0.8 - 1.2\ Å$ outside of the pentagonal plane, and less than $0.4\ Å$ outside of the bidentate plane, with a $\varphi < 30$ deg [35, 92]. The inferior vertex (X) is occupied by a side-chain oxygen, and the superior vertex (-X) is frequently provided by a water molecule. A coordination number of 6-8 is most commonly observed, although coordination numbers ranging from 4-12 have been reported citeGlus1991. Additional structural stability is achieved through hydrogen bonding of the non-ligand Glu and Asp oxygens. The ionic radius of the calcium ion increases with increasing coordination number, but has been generally reported between 0.99 to 1.12Å for a coordination number of 6-8 [70]. As might be expected, the presence of negative charge in the binding site is relevant. Several studies have evaluated the charge environment for $Ca^{2+}$-binding sites, indicating higher binding affinities with a net negative charge within $5 - 15\ Å$ of the binding microenvironment [63, 62], and the presence of 3-4 negative charges in the primary coordination shell may provide optimal charge configuration [69, 110].

For canonical EF-Hands, calcium ions bind in a 12 residue central loop, typically utilizing side-chain oxygen ligands from loop positions 1, 3, 5, 9, and 12, as well as a main-chain carbonyl oxygen from position 7. Ligands associated with EF-Hands are typically Asp at position 1, Asp or Asn at position 3, Asp, Ser or Asn at position 5, a water molecule at position 9, and a bidentate Glu at position 12 [69].

Pseudo EF-Hands coordinate the $Ca^{2+}$ ion predominantly with main-chain carbonyl oxygen atoms in a 14 residue loop. Participating ligands may be contiguous or non-contiguous in the sequence.
Using the volume of information available, numerous attempts have been made to predict $Ca^{2+}$-binding sites using computational methods. The capability of predicting or designing protein structure has far-reaching implications, in that it would allow for protein design with predetermined functions.

Currently, prediction of sites using pattern searches to identify sites based on patterns identified in the primary sequence can be performed using Prosite (http://www.expasy.ch/prosite/). Recent work in our lab resulted in calcium binding-site patterns with higher prediction rates than those presented in the Prosite database \[112\], as well as the capability of predicting non-contiguous binding sites by using variable gap-lengths in the search motif. Prediction of sites based only on primary sequence data suffers from the same constraints as attempting to elucidate tertiary structure based on the sequence alone: mainly, the potential complexity resulting from a multitude of potential conformations.

A number of studies have evaluated algorithms designed to predict binding sites based on pre-existing tertiary structure data. Bagley and Altman developed a grid system to evaluate microenvironment properties surrounding a binding site, score the potential site based on preselected physical and chemical characteristics, and statistically compare the site relative to a control non-site \[5\]. Sodhi et al report a method, MetSite, that combines sequence and structure data to identify residues that form the metal-binding sites \[89\]. This method reportedly resulted in mean accuracy of 94.5% with a true-positive rate of 39.2%, and suggested that structural data alone was insufficient for accurate prediction.

Although other prediction methods have been reported in the literature, only four additional methods/analyses are discussed in this work, as they represent the core datasets utilized in our study.

### 4.2 Method

In general, the common feature for evaluating metal-binding sites in known crystal structures relies on the presence of clustered binding ligands: oxygen, nitrogen, and sulfur. In contrast to assertions noted in previous studies, the presence of identifiable oxygen clusters alone has been shown as sufficient for fast identification of calcium-binding sites in calcium-binding proteins \[22\].
The four datasets presented in this work were selected because they were produced in the past decade, and, for those involving computational predictions, they rely on clustering of ligands in the three-dimensional structure as a precursor to applying their prediction algorithms.

Dataset I was reproduced from Nayal and Di Cera [75], who evaluated 62 $\text{Ca}^{2+}$-binding sites from 32 $\text{Ca}^{2+}$-binding proteins using data from PDB files. As with the Sodhi method, the Nayal study suggests that structural data alone cannot characterize the $\text{Ca}^{2+}$-binding site for prediction purposes [73, 75]. The Nayal and Di Cera algorithm also utilizes a grid system to evaluate valences around a grid point selected based on the presence of at least 3 oxygen atoms within a probe radius of 3.4 Å. A prediction rate of 99.7% was reported using this method, based on the prediction of the $\text{Ca}^{2+}$-binding site within 3.5 Å of the actual site, for a cutoff valence value of 1.6. It is important to note with respect to the Nayal study that the clustering of oxygen atoms was integral to selection of potential sites.

Dataset II was reproduced from Pidcock and Moore, who in 2001 conducted a comprehensive statistical analysis of binding sites for both $\text{Ca}^{2+}$ and lanthanide ions, utilizing data provided in the PDB and the Cambridge Structural Database (http://www.ccdc.cam.ac.uk/) [77]. Dataset II from Pidcock and Moore originally contained 44 proteins with 94 calcium-binding sites. (515 fully normalized crystal structures of calcium-binding proteins in PDB from 1994 to 1999 with a resolution between the range 1.0-2.5 Å. This initial dataset was reduced in our study to 44 structures with 60 sites by removing structures whose calcium-binding sites share ligand residues or ligands other than those donated by the protein and water.)

Dataset III in our study was based on work done by Dudev et al, who evaluated the role of second-shell atoms and their interactions with first shell atoms involved as binding ligands in metal-binding sites [27]. Because the second shell contains many of the main chain or peptide backbone atoms and structures, these constituents contribute to metal selectivity by constraining the metal-binding site with respect to ionic radii and coordination geometry. The Dudev study utilized density functional theory/continuum dielectric methods (DFT/CDM) to evaluate the dielectric medium surrounding the metal-binding site, which was determined by first identifying the oxygen, nitrogen and sulfur atoms surrounding the metal-binding site. The original Dataset consisted of 34 proteins with 81 calcium-binding sites, constrained by PDB structures with a resolution less than 3.0 X-ray and NMR structures, and, with only one exception, no more than 30% sequence
homology with the other selected protein structures.

Dataset IV from Yamashita et al. in 2005 contains 20 proteins with 46 calcium-binding sites (The PDB short list containing calcium is used for analyze the relationship between experimental and predicted energy). Yamashita et al reported the use of a radial distribution plot surrounding a metal binding site to develop a hydrophobicity contrast function, which identifies metal binding sites based on the assumption that a hydrophilic core around the binding site is surrounded by spherical hydrophobic shell [107]. The determination of the hydrophilic core within the grid system is determined based on the presence of clustered charged-ligands: essentially, oxygen, nitrogen and sulfur.

The characteristics of the calcium-binding sites from the selected datasets that were analyzed statistically in this study are graphically presented in Figure 4.1.

As see in Figure 4.1, Ca-O-C is the angle between the central $Ca^{2+}$ (Ca) ion, the ligand oxygen (O), and the carbon (C) bound to the ligand oxygen. O-Ca-O is the angle between two oxygen ligands (O) and the central $Ca^{2+}$ (Ca) ion. Distances evaluated were a) between the central $Ca^{2+}$ (Ca) and and the carbon (C) bound to the ligand oxygen ($dist(Ca,C)$), and b) between the ligand oxygen (O), and the carbon (C) bound to the ligand oxygen ($dist(C,O)$). Distinction is also made in the statistical analysis between monodentate and bidentate ligands.

Angle and distance values were calculated using Matlab7.0. Statistics were compiled and graphically-rendered using MS Excel (Microsoft corporation, Redmond, Wa)

### 4.3 Results

Table 4.1 summarizes the coordination numbers (CN) for the calcium binding sites evaluated in each of the four sample datasets. For the purpose of this study, a cutoff distance of 3.5Å was selected, so that only ligand atoms (O, N) within this distance would be considered coordinating ligands. As a consequence of this distance constraint, the CN values reported in table differ from those reported for the original datasets. The majority of coordination number (CN) is 6, 7, and 8. 3 14% of the datasets are complexes with a CN of 3 or 4. The average CN per calcium-binding site is from 6.4 to 6.9.

It should be noted that ligands can be from the protein itself, cofactors, or water molecules. Ta-
Figure 4.1: Illustration of two angles: Ca-O-C and O-Ca-O, and two distances: dist(Ca,C) and dist(C,O)
Table 4.1: Coordination numbers (CN) of four datasets

<table>
<thead>
<tr>
<th>Dataset</th>
<th>Total Sites</th>
<th>CN8</th>
<th>CN7</th>
<th>CN6</th>
<th>CN5</th>
<th>CN4</th>
<th>CN3</th>
<th>Avg CN</th>
</tr>
</thead>
<tbody>
<tr>
<td>I</td>
<td>62</td>
<td>10</td>
<td>31</td>
<td>14</td>
<td>4</td>
<td>0</td>
<td>2</td>
<td>6.7</td>
</tr>
<tr>
<td>II</td>
<td>94</td>
<td>11</td>
<td>35</td>
<td>32</td>
<td>11</td>
<td>4</td>
<td>1</td>
<td>6.4</td>
</tr>
<tr>
<td>III</td>
<td>81</td>
<td>14</td>
<td>37</td>
<td>19</td>
<td>1</td>
<td>6</td>
<td>4</td>
<td>6.4</td>
</tr>
<tr>
<td>IV</td>
<td>46</td>
<td>7</td>
<td>32</td>
<td>3</td>
<td>3</td>
<td>0</td>
<td>1</td>
<td>6.9</td>
</tr>
</tbody>
</table>

Table 4.2 summarizes the sources of the ligand atoms involved in binding site coordination. Side-chain carboxylate oxygen atoms represent the major ligand contributor, followed by water molecules, which stresses the significance of solvent exposure during binding.

4.3.1 Average Ca-O and Average Ca-C

Figure 4.2 presents the average distance values for ligand oxygen atoms (first shell) and their associated carbon atoms (second shell). The Avg Ca-O and Avg Ca-C values were calculated as follows in Equation 4.1 and Equation 4.2, respectively.

\[
RO = \frac{\sum_{i=1}^{k} dist(Ca, O)}{k} \tag{4.1}
\]

\[
RC = \frac{\sum_{i=1}^{k} dist(Ca, C)}{k} \tag{4.2}
\]

In Equation 4.1, k is the number of ligands in one site. In Equation 4.2, m is the number of bonded carbon atoms, and \( k \geq m \). When k equals m in a single binding site, it indicates that only monodentate ligands appear in this site, otherwise k must be greater than m for polydentate ligands. From figure 2., it can be observed that the peak value of Avg Ca-O is either 2.4 (dataset I) or 2.5 (dataset II, III, IV), which in fact means Avg Ca-O per site is between 2.3 and 2.5 because we use 0.1 as interval bin. Also, the peak values of Avg Ca-C are 3.3 for all four datasets, which means Avg Ca-C per site is between 3.2 and 3.3 due to bin interval. For all four datasets, the resulting curves are very similar, and exhibit a characteristic decrease as the distance for both the ligand oxygens and their associated carbon atoms approaches the first/second shell interval, which is expected.
4.3.2 $r_{RO\_RC}$ and $ar_{RO\_RC}$

To examine the spatial relationship between the inner oxygen shell and the outer bonded carbon shell of calcium-binding sites, the ratio of Avg Ca-O and Avg Ca-C ($r_{RO\_RC}$), is calculated using Equation 4.3. It is hypothesized that a fixed ratio should exist for this relationship, based on the average C-Ca and average O-Ca bond lengths and constraints on the the C-O-Ca angle. Equation 4 calculates an adjusted ratio, where NB represents the number of bidentate residue(s) in a putative calcium-binding site and 0.5 is an empirically-derived coefficient of NB to determine the deduction effect on $ar_{RO\_RC}$

The calculated ratios $r_{RO\_RC}$ and $ar_{RO\_RC}$ for the four datasets are represented in Figure 4.3a and Figure 4.3b, respectively.

$$r_{RO\_RC} = \frac{RO}{RC}$$  \hspace{1cm} (4.3)

$$ar_{RO\_RC} = r_{RO\_RC} - 0.5NB$$  \hspace{1cm} (4.4)
Figure 4.3: The distributions of $r_{RO \cdot RC}$ (a) and $ar_{RO \cdot RC}$ (b) in dataset I (light gray), dataset II (dark gray), dataset III (white), and dataset IV (black)

It is clear that the majority of $r_{RO \cdot RC}$ is represented by the value ranging from 0.66 to 0.84 with an interval bin of 0.03. After adjustment of $r_{RO \cdot RC}$, the majority of $ar_{RO \cdot RC}$ is represented by the value ranging from 0.66 to 0.78. Therefore, a relatively stable range value exists to quantitatively evaluate the geometric properties of calcium binding sites. Besides calcium-binding sites, are there any other domains or locations of proteins which have similar geometric property? It is an open question motivating us to explore in the prediction of calcium binding sites in proteins using this measurement of geometric property without knowing documented calcium location.

4.3.3 Bidentate Residue Effect on $r_{RO \cdot RC}$

More interestingly, we find that the bidentate property of both aspartate and glutamate have an important impact on the distribution of $r_{RO \cdot RC}$. Intuitively, the carbon shell will get closer to the oxygen shell in one calcium-binding site if there is a bidentate residue in that site, as the carbon from the bidentate residue is restricted to stretch away from the calcium cation. The effect of differentiating sites based on the presence of bidentate ligands was examined for all four datasets. Similar results were obtained for all analyses. The results for only Dataset II, which included the most proteins of the four evaluated datasets, are depicted in Figure 4.4. As seen in Figure 4.4, changes in $r_{RO \cdot RC}$ based on inclusion of bidentate ligands results in a decreased ratio, which is to be expected, as the bidentate ligand distances (O-Ca) are shorter than non-bidentate distances.
Figure 4.4: Distribution of $r_{RO\_RC}$ for non-bidentate (black) and bidentate sites (white)

A clear cutoff between the two distributions is apparent at 0.75. Even though there is the overlap between 0.72 and 0.75 of $ar\_RO\_RC$ for both sites, the frequency (3 sites) from bidentate sites is very small. While Figure 4 illustrates the distinction between sites including bidentate ligands, it does not explain this phenomenon. To answer that, it is necessary to evaluate the Ca-O-C bond angle.

4.3.4 Ca-O-C

The Ca-O-C angle constrains the proximity of carbon from the calcium cation to some extent. A decrease in this angle generally results in a decrease in the C-Ca distance. It can be seen in Figure 5 that the bidentate Ca-O-C angles range between 100 and 150 whereas the monodentate (i.e. - non-bidentate) Ca-O-C angles range between 130 and 160. These ranges overlap between 130 and 150 which also corresponds with the highest frequency distribution across the datasets. The
Figure 4.5: Distribution of Ca-O-C angles for bidentate and monodentate residues over four datasets
bidentate angle properties can also be characterized by the sum of the angle pairs, as seen in Figure 4.6. Here we see a more discernable quantity, where the majority of the frequency distribution values fall between 180 and 200, with a bin interval of 10.

4.3.5 Relationship Between $r_{RO_{RC}}$ and Ca-O-C

In order to relate the individual ligand Ca-O-C angles to $r_{RO_{RC}}$ for the binding site, the average Ca-O-C (denoted as AvgCa-O-C) angle is defined by calculating the mean for all Ca-O-C angles in each individual site. In this way, both AvgCa-O-C and $r_{RO_{RC}}$ can be compared. Figure 4.7 shows the correlation of AvgCa-O-C and $r_{RO_{RC}}$ for Dataset II.

As seen in Figure 4.7, the coordinate convergence appears linear, with AvgCa-O-C decreasing in inverse proportion to $r_{RO_{RC}}$. It can also be observed that AvgCa-O-C is always greater than 90, which implies that the carbon atom must reside at a greater distance than the oxygen ligand from the calcium cation. To verify the linear relationship between these variables, a Pearson coefficient
Figure 4.7: Distribution of Ca-O-C angle pair sums for bidentate residues over four datasets

was calculated at -0.96 which means that strong negative correlation exists. The Pearson product-
moment correlation coefficient (r) is detailed in Equation 4.5.

$$r = \frac{\sum(x - \bar{x})(y - \bar{y})}{\sqrt{\sum(x - \bar{x})^2(y - \bar{y})^2}}$$  \hspace{1cm} (4.5)

In Equation 4.5, (r) is a measure of the correlation of two variables X and Y measured on
the same object or organism, that is, a measure of the tendency of the variables to increase or
decrease together. The values for x bar and y bar are the sample means AVERAGE(x) and
AVERAGE(y). The value r ranges from -1 to 1. A value of 1 indicates that a linear equation
describes the relationship perfectly and positively, with all data points lying on the same line and
with Y increasing with X. A score of -1 shows that all data points lie on a single line but that Y
increases as X decreases. A value of 0 shows that a linear model is inappropriate - that there is no
linear relationship between the variables.
4.3.6 Types of Ligands per Site

Because $Ca^{2+}$ is a positively-charged ion, the microenvironment surrounding the ion must necessarily exhibit a net negative charge of -2. It was previously noted that 3-4 negative charges in the primary coordination shell may provide optimal charge configuration [69, 110]. In the studies cited for datasets used in this work, formal charge is defined where sidechain carboxyl groups have a charge of -1, main chain carbonyls from Glu or Asp have a charge of -1, and all other main chain carbonyls have a formal charge of zero. However, data summarized in the four datasets used in this study indicate the frequency of charges follows the order $-3 > -2 > -1 > -4$, as seen in Figure 4.8.

Interestingly, six sites were found with a formal charge of zero, and six sites with a formal charge of 5. To further examine this, data from PDB files for 1BJR and 1HYT were obtained. Relevant data from the PDB files is summarized in Table 4.3. The columns in Table 4.3 are defined as follows:
HetAtm Res indicates the sequence number associated with the calcium ion; ResName is the amino acid providing the ligand atom; AtomCharID indicates the atom type; ResSeq is the primary sequence number of the residue; Ca Bind Dist is the distance between the ion and the ligand atom; C-Lig-Ca Angle is the angle between the calcium ion, the binding ligand atom and its associated carbon; Ligand source indicates whether the ligand atom is from the main chain or the side chain, and the functional group; and Charge indicates the formal charge associated with each ligand atom.

Models of the two binding sites were completed using Pymol (http://pymol.sourceforge.net/), and are presented in Figure 4.9.

As seen in Table 4.3, the 4 binding ligands surrounding calcium 290 in 1BJR are all carbonyl oxygens, with no formal charge. For calcium 804 in 1HYT, seven ligand atoms bind the calcium ion, comprised of two pairs of bidentate ligands that each contribute a charge of -1, two side chain carboxyl groups that each contribute a charge of -1, and a final charge of -1 from a main chain Glu carbonyl.

Figure 4.9 indicates an irregular geometry for both binding sites, as compared with the more symmetric geometry typically seen in EF-Hand calcium binding motifs (Figure 4.10). Nonetheless, the structures observed in Figure 4.9 agree with the calculated charges
Figure 4.9: A) 1BJR calcium 290 has a formal charge of zero in the binding site, and B) 1HYT, calcium 801, has a formal charge of 5
Table 4.2: Ligand Charge Data for Two Calcium Binding Sites

<table>
<thead>
<tr>
<th>PDB_ID</th>
<th>HetAtm</th>
<th>ResName</th>
<th>Atom</th>
<th>CharID</th>
<th>ResSeq</th>
<th>Dist</th>
<th>Angle</th>
<th>Ligand source</th>
<th>Charge</th>
</tr>
</thead>
<tbody>
<tr>
<td>1BJR</td>
<td>290</td>
<td>Arg</td>
<td>O</td>
<td>12</td>
<td>2.77</td>
<td>149.18</td>
<td>MC Carbonyl</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td>1BJR</td>
<td>290</td>
<td>Ser</td>
<td>O</td>
<td>15</td>
<td>3.28</td>
<td>169.84</td>
<td>MC Carbonyl</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td>1BJR</td>
<td>290</td>
<td>Asn</td>
<td>OD1</td>
<td>257</td>
<td>2.83</td>
<td>144.17</td>
<td>SC Carbonyl</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td>1BJR</td>
<td>290</td>
<td>Ala</td>
<td>O</td>
<td>273</td>
<td>2.98</td>
<td>132.93</td>
<td>MC Carbonyl</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td>1HYT</td>
<td>801</td>
<td>Asp</td>
<td>OD2</td>
<td>138</td>
<td>2.32</td>
<td>132.93</td>
<td>SC Carboxyl</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td>1HYT</td>
<td>801</td>
<td>Glu</td>
<td>OE1</td>
<td>177</td>
<td>2.45</td>
<td>101.84</td>
<td>SC Carboxyl</td>
<td>0.5</td>
<td></td>
</tr>
<tr>
<td>1HYT</td>
<td>801</td>
<td>Glu</td>
<td>OE2</td>
<td>177</td>
<td>2.82</td>
<td>82.44</td>
<td>SC Carboxyl</td>
<td>0.5</td>
<td></td>
</tr>
<tr>
<td>1HYT</td>
<td>801</td>
<td>Asp</td>
<td>OD1</td>
<td>185</td>
<td>2.43</td>
<td>132.09</td>
<td>SC Carboxyl</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td>1HYT</td>
<td>801</td>
<td>Glu</td>
<td>O</td>
<td>187</td>
<td>2.27</td>
<td>150.98</td>
<td>MC Carbonyl</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td>1HYT</td>
<td>801</td>
<td>Glu</td>
<td>OE1</td>
<td>190</td>
<td>2.48</td>
<td>92.90</td>
<td>SC Carboxyl</td>
<td>0.5</td>
<td></td>
</tr>
<tr>
<td>1HYT</td>
<td>801</td>
<td>Glu</td>
<td>OE2</td>
<td>190</td>
<td>2.51</td>
<td>89.81</td>
<td>SC Carboxyl</td>
<td>0.5</td>
<td></td>
</tr>
</tbody>
</table>
Figure 4.10: Pentagonal-bipyrimid geometry surrounding $Ca^{2+}$ ion (A) and (B) ligand-ion contact geometry for calmodulin EF-I binding pocket from PDB 3CLN. The water molecule at -X is not shown.
Chapter 5

PREDICTING DISULFIDE CONNECTIVITY

5.1 Introduction

Disulfide bonds are covalent bonds between two non-adjacent cysteine residues in proteins. They play an essential role in folding and assembling of proteins [41, 86]. The better prediction on disulfide bonds will lead the better prediction on protein structure. Furthermore, the knowledge of disulfide bonds helps reveal structure/function relationship of proteins. Therefore, developing methodology and/or descriptors for the prediction of disulfide connectivity revealing properties of the bonds is as important as the gaining the high accuracy.

Although there have been bunches of approaches using different techniques, Tsai et al. have categorized all methods into two groups: (1) pattern-wise as shown in [101, 111] and (2) pair-wise as shown in [32, 7, 31]. The idea of a pattern-wise method is based on comparing the test protein sequence with template proteins having the same number of disulfide bonds and assign the connectivity pattern of the template protein, whose primary sequence is most similar in some aspect to that of the test protein; the basic principle of a pair-wise method is to break the test protein into all possible bonds, then associate a probability function for each possible bridge by comparing them with all bridges in a template dataset, finally assign the set of disjoint pairs with maximum total weight to the disulfide bond connectivity, which usually carries out by applying the well-known maximum weight perfect matching algorithm developed by Edmonds [28].

The highest prediction accuracy among all results quoted above is 52% obtained by [111]. Recently, elegantly applying the state-art technology SVM, Tsai et al. improved the accuracy to 63%; Chen et al. established a new record of prediction accuracy of 70% by combining the previous works [97, 111] into a two level model in addition to using SVM [20]. after calculating the bond possibility of any pair of cysteines in one protein chain; consequently, they belong to pair-wise method in general. The main reason why the accuracy of pattern-wise method is low is mainly
due to there do not have sufficient template proteins having the same number of disulfide bonds as test proteins. Notice that the result in the work [20] exceeds the best performed pattern-wise method [111] by 18% in prediction accuracy, which is a giant leap. Zhao et al. and Tsai et al. have realized that the limitation comes from the number of template patterns in pattern-wise methods [111, 97]. In order to break this limitation, we develop a new method to compare proteins with different number of disulfide bonds if there is no protein with the same number of disulfide bonds close enough to a test protein.

A descriptor is another determinant factor for prediction performance. Fariselli and Casadio calculated the bond probability of two cysteines from local contact potential profile [31]. Zhao et al. used global sequence separations as the representation of a protein [111]. Tsai et al. also used global sequence separations as one of their descriptors for predicting the bond probabilities of possible pairs of cysteines [97]. Secondary structure has been employed as the descriptor for input coding [7, 32]. Inspired from previous descriptors, we construct a global descriptor of secondary structure, cysteine separation profile of secondary structure (CSPSS), similar to the CSP descriptor [111]. Instead of using a local window of secondary structures, we extend the window beginning from the first cysteine and ending to the last cysteine in one protein chain.

Our approach combines the new proposed pattern-wise method and the descriptor of CSPSS to predict disulfide connectivity. We apply the method on SWISS-PROT 39 (SP39) [6] and obtain the prediction accuracy of 70%.

5.2 Method

The new prediction consists of three steps. First, the secondary structures are generated from protein sequences by applying the PSIPRED server of [71]. Then, the CSPSSs are encoded on the separation profile of cysteine positions on the secondary structure sequence. Finally, the new pattern-wise method is used to assign known patterns to test proteins according to the divergences of CSPSSs between the test proteins and template proteins.
5.2.1 CSPSS

Let $x$ be a protein with $n$ disulfide bonds paired by $2n$ cysteine residues. The cysteine separation profile on secondary structure ($CSPSS^x$) is defined as follows,

$$CSPSS^x = (S_1, S_2, ..., S_{2n-1})$$

(5.1)

where $S_i$ represents the segment of secondary structure sequence between the $i$th paired cysteine and the $(i+1)$th paired cysteine for each $i = 1, 2, ..., 2n - 1$. By the definition, $S_i$ is a character array consisting of three possible symbols (C, H, E) representing different secondary structures. To make it numeric, we transform $S_i$ into $\hat{S}_i$, a vector of coordinates as follows,

$$\hat{S}_i = (L_i, C_i, E_i, H_i),$$

(5.2)

where $L_i$ is the length of the segment, $C_i$ is the number of C’s, $E_i$ is the number of E’s, and $H_i$ is the number of H’s. Clearly, $L_i = C_i + E_i + H_i$. The following is an example showing how to form $S_i$ and $\hat{S}_i$ from a given protein chain (AMCI_APIME) containing 56 amino acids (aa), 10 paired cysteines, the disulfide pattern [1−7, 2−5, 3−6, 4−10, 8−9], and 56 secondary structures (ss) from prediction.

\[\text{aa} : \quad EECGPNEFNTCGSACAPTCAQPKTRIC\]
\[\text{ss} : \quadCCCCCCCCCCCCCCCCCCCCCCCC\]
\[S_1 = [CCCCCC] \rightarrow \hat{S}_1 = [10, 10, 0, 0] \]
\[S_2 = [CCCC] \rightarrow \hat{S}_2 = [5, 5, 0, 0] \]
\[\vdots \]
\[S_8 = [CCCCCC] \rightarrow \hat{S}_8 = [13, 11, 2, 0] \]
\[S_9 = [CCCCCC] \rightarrow \hat{S}_9 = [7, 7, 0, 0] \]
5.2.2 An expanded pattern-wise method

The CSPSS of a test protein could be compared with all CSPSS’s of template proteins. The disulfide connectivity pattern of the test protein can be predicted as that of the template protein with the most similar CSPSS, i.e. with the minimum divergence value. The divergence between protein $x$ and protein $y$, $D(x, y)$, is defined below:

$$|\hat{S}_i^x - \hat{S}_i^y| = (L_i^x - L_i^y)^2 + (C_i^x - C_i^y)^2 + (E_i^x - E_i^y)^2 + (H_i^x - H_i^y)^2,$$

and

$$D(x, y) = \sum_{i=1}^{2n-1} \sqrt{|\hat{S}_i^x - \hat{S}_i^y|},$$

where $S_i^x$ and $S_i^y$ are the $i$th separations of the CSPSSs of two proteins of $x$ and $y$, respectively. Similarly, $\hat{S}_i^x$ and $\hat{S}_i^y$ are the $i$th vectors of the CSPSSs; $L_i^x$ and $L_i^y$ are the length of $i$th segments; $C_i^x$ and $C_i^y$ are the number of C’s in $i$th segments; $E_i^x$ and $E_i^y$ are the number of E’s in $i$th segments; and $H_i^x$ and $H_i^y$ are the number of H’s in $i$th segments of different proteins according to the letter in the following parenthesis. If more than one proteins meet minimum value for one test protein, one of the template patterns will be selected randomly. In fact, this situation has not occurred in our experiments.

The known pattern-wise method algorithms [101, 111] restrict comparison only between a test protein and a template protein with the same number of disulfide bonds. This restriction may reach the limitation when these proteins with the same number of disulfide bonds are not very close to the test proteins, i.e., the divergencies may not be small enough. To overcome the obstacle, we expand the pattern-wise method to compare test proteins with template proteins having different number of disulfide bonds. The new method includes four possible phases described as follows for predicting the disulfide bond pattern of a given test protein $x$. For any protein $p$, $B^p$ denotes the set of all disulfide bonds and $|B^p|$ denotes the number of disulfide bonds in $p$.

1. Calculate $D(x, y)$ for all template proteins $y$ satisfying that $|B^x| = |B^y|$; assign the disulfide bound pattern of the $y$ to the pattern of $x$ if there is a $y$ such that $D(x, y)$ is minimum one and is less than the pre-determined cutoff value. Otherwise,
2. Calculate $D(x, y)$ for all artificially modified template proteins $y$ satisfying that $|B^x| = |B^y|$ where $y$ is modified from $z$ by removing one or more disulfide bonds in $z$ ($y$ has same primary sequence and secondary structure with $z$, and $B^y \subset B^z$); assign the pattern of $y$ to the pattern of $x$ if there is a $z$ with its modified protein $y$ such that $D(x, y)$ is minimum one and is less than the cutoff value. Otherwise,

3. Modify $x$ into two proteins: $x_i$ and $x_j$ by removing one or more disulfide bonds in $x$ where $x_i$ and $x_j$ have to satisfy the following conditions: $B^x = B^{x_i} \cup B^{x_j}$, $B^{x_i} \cap B^{x_j} = \emptyset$, and $|B^{x_i}| > |B^{x_j}|$; compare $x_i$ with any template protein $u$ satisfying that $|B^u| = |B^{x_i}|$ and $x_j$ with any template protein $v$ satisfying that $|B^v| = |B^{x_j}|$; assign the pattern of $u$ to the pattern of $x_i$ if there is a $u$ such that $D(x_i, u)$ is minimum one and is less than a cutoff value, and the pattern of $v$ to the pattern of $x_j$ if there is a $v$ such that $D(x_j, v)$ is minimum one and is less than a cutoff value; assemble the patterns of $x_i$ and $x_j$ into the candidate pattern of $x$ (denoted as $P(|B^{x_i}| + |B^{x_j}|)$ if they are available; choose one of candidate patterns with the largest value of $|B^{x_i}|$ if there are more than one candidate patterns.

4. If no such satisfied pattern of $x$ is assembled in the third phase, take the pattern with minimum divergence without cutoff restriction from all above phases. As a matter of fact, the minimum divergence in the 3rd phase is the sum of both minimum divergences through comparisons between two modified proteins of $x$ and template proteins.

It is worth mentioning three points detailed as follows in the implementation of the method.

$p(B)$ denotes the protein $p$ with $B$ disulfide bonds.

1. The template proteins with relatively small number of disulfide bonds have higher priority for comparison than those with relative larger number of disulfide bonds in the 2nd phase. For instance, if $x(2)$ can find a satisfied pattern from a template protein $y(3)$, $x(2)$ will not search pattern in any template protein $y(B)$ where $B > 3$.

2. Pattern assembly has a preference order in the 3rd phase. For instance, $x(5)$ does not have any match in the first two phases. $x_i(4)$ can find a match from $u(4)$ in the 3rd phase, the pattern of $u(4)$ adds the left pair of cysteines ($v(1)$, a special case) to assembly the candidate
pattern of \(x(5), P(4 + 1)\). Similarly, \(x\) can be divided into two modified proteins of \(x_i(3)\) and \(x_j(2)\), and can obtain a candidate pattern of \(P(3 + 2)\) assembled from the matched patterns of \(u(3)\) and \(v(2)\). Instead of the \(P(3 + 2)\), the \(P(4 + 1)\) should be assigned to the pattern of \(x(5)\) because of the preference order. In fact, pattern-searching stops immediately after finding the \(P(4 + 1)\) in the implementation.

3. The empirical value of \(6B\) is taken as the cutoff value of divergences where \(B\) is the number of disulfide bonds.

5.2.3 Measurement

The prediction accuracy of our method is also evaluated with \(Q_P\), which is the fraction of proteins with correct disulfide connectivity pattern prediction and is defined as:

\[
Q_P = \frac{C_P}{T_P}
\]  

(5.3)

where \(C_P\) is the number of proteins correctly predicted and \(T_P\) is the number of total test proteins. We ignore another general measurement of the fraction of pairs of cysteines correctly predicted in our experiments because even one displaced disulfide bond in a protein may lead to a nonfunctional protein structure or a protein structure with different function \([37, 36, 86]\). The \(Q_P\) is more reliable to evaluate performance on disulfide bond prediction.

5.3 Results

5.3.1 Dataset

To compare our method to all other methods published in 2005-2006, the dataset from SWISS-PROT named SP39 \([6]\) was adopted for method validation. To avoid the influence of sequence homology, the dataset was divided into four groups to guarantee that each two proteins from different groups have a sequence identity less than 30\% \([7, 97]\). The numbers of sequences according to the bonds are displayed in Table 5.1. as well as the results.
Table 5.1: Performance comparison among different algorithms

<table>
<thead>
<tr>
<th>Method</th>
<th>B=2</th>
<th>B=3</th>
<th>B=4</th>
<th>B=5</th>
<th>B=25</th>
</tr>
</thead>
<tbody>
<tr>
<td>SP39</td>
<td>156</td>
<td>146</td>
<td>99</td>
<td>45</td>
<td>446</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Method</th>
<th>Q_P</th>
<th>Q_P</th>
<th>Q_P</th>
<th>Q_P</th>
<th>Q_P</th>
</tr>
</thead>
<tbody>
<tr>
<td>2D – RNN&lt;sup&gt;a&lt;/sup&gt;</td>
<td>74</td>
<td>51</td>
<td>27</td>
<td>11</td>
<td>49</td>
</tr>
<tr>
<td>DiANNA&lt;sup&gt;b&lt;/sup&gt;</td>
<td>62</td>
<td>40</td>
<td>55</td>
<td>26</td>
<td>49</td>
</tr>
<tr>
<td>CSP&lt;sup&gt;c&lt;/sup&gt;</td>
<td>72</td>
<td>54</td>
<td>33</td>
<td>18</td>
<td>52</td>
</tr>
<tr>
<td>CSPSS+Old Pattern-wise</td>
<td>72</td>
<td>58</td>
<td>37</td>
<td>18</td>
<td>54</td>
</tr>
<tr>
<td>SVM&lt;sup&gt;d&lt;/sup&gt;</td>
<td>79</td>
<td>53</td>
<td>55</td>
<td>71</td>
<td>63</td>
</tr>
<tr>
<td>Two - levelModel&lt;sup&gt;e&lt;/sup&gt;</td>
<td>85</td>
<td>67</td>
<td>57</td>
<td>58</td>
<td>70</td>
</tr>
</tbody>
</table>

CSPSS+New Pattern-wise       | 79  | 68  | 68  | 45  | 70  |

<sup>a</sup> [7]  
<sup>b</sup> [32]  
<sup>c</sup> [111]  
<sup>d</sup> [97]  
<sup>e</sup> [20]
5.3.2 Cross-validation

In order to compare with other methods for disulfide connectivity prediction, same criteria were applied on selecting our dataset. Also the same fourfold cross-validation has been applied on our dataset. Even the selection of four subsets is same as the method [7]. The SP39 were divided into four subsets each of which has four balanced groups according to the number of bonds. Table 5.1 lists the accuracies of four-fold cross-validation performed with the dataset SP39 for our method along with some results reported previously. Here we only list the results published in 2005 because the results before 2005 show the accuracy is up to 46%. Baldi et al. used 2-Dimensional Recursive Neural Network (2D-RNN) to predict disulfide connectivity in proteins starting from their primary sequence and its homologues [7]. The outputs of 2D-RNN are the pair-wise probabilities of the existence of a bridge between any pair of cysteines. Finally, the weighted matching algorithm is applied on the graph with all edges/possibilities between any two vertices/cysteines. A diresidue Neural Network (DiANNA) [32] is trained to recognize pairs of bonded half-cystines given input of half-cystines symmetric flanking regions. The network is trained using disulfide bonds information derived from high-quality protein structures. the data are encoded with respect to cysteine pairs. Zhao et al. simply adopted the descriptor of cysteine separation profile and used the nearest neighboring method [111]. For the SVM model [97], the features coded are the information extracted from profile and distances between oxidized cysteines (DOC). After the data are encoded, the SVM model is used to predict bonding probabilities for each cysteine pair. Finally, the problem is transformed into a maximum weight matching problem and solved to find the final bonding pattern for a protein.

With the CSPSS descriptor the traditional pattern-wise method obtained a Qp of 54%, which is better than those obtained in previous methods except the SVM models. The reason for the improvement is the consideration of global secondary structure. We also found that our results contain all but one correct predictions from the CSP method [111], which validates that our descriptor includes the information of the CSP descriptor. Furthermore, we combine the CSPSS descriptor with the new pattern-wise method to obtain the best performance of 70%, which is comparable to the current-of-the-state method, two-level SVM model [20]. More interestingly, our method reflects the evolutionary process of DNA recombination and mutation between the
structure patterns with fewer disulfide bonds and the structure patterns with more disulfide bonds. Figure 5.1 shows an example of finding pattern in the template protein with more disulfide bonds through the new pattern-wise method. ANGI_HUMAN (PDBid:1A4Y, pattern:[1-4, 2-5, 3-6], Figure 5.1a) can not find a correct pattern from template proteins with three disulfide bonds using the traditional pattern-wise method [111] because even the minimum divergence between the test protein and template proteins is too large. However, it can find the correct pattern from the modified protein with the pattern of [1-4, 2-5, 3-6](26-84, 40-95, 58-110) from RNP_RAT (PDBid:1A4Y, pattern:[1-6, 2-7, 3-8, 4-5](26-84, 40-95, 58-110, 65-72) Figure 5.1b) by removing the bond of 4-5(65-72) through the 2nd phase of the new pattern-wise method. It indicates that the protein with more disulfide bonds may evolve from the protein with fewer disulfide bonds. Figure 5.2 shows an example of finding pattern in the template protein with fewer disulfide bonds through the new pattern-wise method, which may add proof to the above opinion. UROK_HUMAN (PDBid:1LMW, pattern:[1-3, 2-4, 5-9, 6-7, 8-10], Figure 5.2a) can not find a correct pattern from the template protein of five disulfide bonds using the traditional pattern-wise method [111], but it can be divided into two modified proteins one of which has paired cysteins (42, 58, 136, 168, 182, 191, 201, 220), then this modified protein can find the satisfied pattern from EL1_PIG (PDBid:1C1M, pattern:[1-2, 3-5, 4-7, 6-8](42-58, 136-201, 168-182, 191-220) Figure 5.2b), finally the pattern with paired cysteines adds the left pair [1-2](50-111) in the other modified protein to make a complete pattern according to the index of paired cysteines in protein 1LMW through the 3rd phase of the new pattern-wise method.

5.4 Conclusion and Discussion

There are two major categories for the descriptors of disulfide connectivity prediction: (1) The global descriptors [20, 97, 111] such as sequence length and the positions of all cysteines, and (2) local descriptors [7, 20, 31, 32, 97, 101] such as secondary structure and residue contact potential. The SVM model [97] for predicting disulfide connectivity benefits from the combination of the global descriptor of distance of cysteines (DOC) and local descriptor (sequence profile). The two level SVM model [?] takes advantages of both the CSP descriptor [?] and the descriptors from previous SVM model [97]. We add the descriptor CSPSS to the global descriptors, test this
Figure 5.1: An example of finding the pattern of the test protein (a) from the subpattern of the template protein (b) with more disulfide bonds.

1A4Y: 26-81 39-92 57-107
1RRA: 26-84 40-95 58-110 65-72
Figure 5.2: An example of finding the pattern of the test protein (a) from the template protein (b) with less disulfide bonds.
descriptor with traditional and new pattern-wise methods, and obtain good performance. It is thus possible to improve the results through combining the CSPSS and local descriptors.

There are two major categories for the methods of disulfide connectivity prediction: pattern-wise and pair-wise already mentioned in the section of introduction. We develop the new pattern-wise method which can replace the traditional pattern-wise method because of the limited number of patterns available in the template set of proteins. The new pattern-wise method could also replace pair-wise methods to some extent not only because it is simple and time-efficient but also because it considers the whole pattern naturally not breaking any disulfide bond. In the literature [97], they mentioned that each of CTRA_BOVIN (PDBid:1HJA, pattern: [1-4, 2-3, 5-9, 6-7, 8-10]) and UROK_HUMAN (PDBid:1LMW, pattern:[1-3, 2-4, 5-9, 6-7, 8-10], Figure 5.2a) has an unique pattern in the dataset SP39 and can not be predicted by the pattern-wise method CSP [111]. Our new pattern-wise method can successfully predict the disulfide patterns of both proteins as well as the SVM model [97]. In addition, there exists some deficiency of machine learning techniques such as neural networks and SVM that they always take the non-bonded cysteines as negative samples and are thus not scientific without considering influences between disulfide bonds sufficiently in one protein chain. The method itself could be refined through the process of dividing test proteins in the 3rd phase. Meanwhile, the performance of the method could be improved by balancing the tradeoff between divergence cutoff value and the hierarchical priority through more experiments of the method. In this study, we propose a new pattern-wise method to resolve disulfide pattern insufficiency which the traditional pattern-wise method causes. This method takes full advantages of pattern analysis considering the influences among disulfide bonds in one protein chain. The best performance reached by this method with the descriptor of CSPSS indicates that the simple pattern-wise method is comparable to the best pair-wise method and protein secondary structure is highly related to disulfide bonds. Both the method and the descriptor of CSPSS are available for use with other methods and descriptors for predicting disulfide connectivity.
Chapter 6

CONCLUSIONS

This dissertation work focuses on two bioinformatics problems: calcium-binding and disulfide bond connectivity. I have tried so many graph theory algorithms to solve these practical issues. Because these two problems involve bunches of biophysical, biochemical, geometrical properties, some algorithms may only partially solve the problems and provide a framework or pathway to reveal the interior relationship among certain properties. All work done are from computational angle and are expected to verification.
BIBLIOGRAPHY


